Polskie Towarzystwo Semiotyczne

STUDIA SEMIOTYCZNE— ENGLISH SUPPLEMENT

Volume XXX

Warszawa • 2019

Adres redakcji: Krakowskie Przedmieście 3, 00-047 Warszawa
e-mail: studiasemiotyczne@pts.edu.pl
http://studiasemiotyczne.pts.edu.pl/

ISSN 0137-6608; e-ISSN 2544-073X

Prace redakcyjne związane z przygotowaniem składu elektronicznego, korektą językową i zamieszczaniem artykułów naukowych na stronie internetowej „Studiów Semiotycznych" i na stronie internetowej „Studiów Semiotycznych—English Supplement" – zadanie finansowane w ramach umowy 636/P-DUN/2019 ze środków Ministra Nauki i Szkolnictwa Wyższego przeznaczonych na działalność upowszechniającą naukę.

Ministerstwo Nauki
i Szkolnictwa Wyższego

CONTENT

Studia Semiotyczne—English Supplement, t. XXX (2019), s. 5–19
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/stsen.xxx.01
Article

Piotr Wilkin *

NATURALIZED REPRESENTATIONS— A USEFUL GOAL OR A USEFUL FICTION?

Summary: One of the key concepts of naturalized epistemology as well as the cognitive sciences that stem from it is the naturalized concept of mental representation. Within this naturalized concept, many attempts have been made to unify (for humans as well as for other living organisms) the notion of representation error. This text makes an attempt to argue against the adequacy of using a naturalized concept of representation error as well as casts doubt on the wide program of naturalizing concepts related to human conceptuality.

Keywords: mental representations, representation error, naturalization.

1. Introduction

In philosophy of mind, the naturalistic approach is becoming more and more popular; it is also a constitutive approach, if not for cognitive science, then at least for some branches of it. One of the fundamental concepts of cognitive science which is often naturalized is the concept of *cognitive representation*. One of the most popular approaches to naturalizing

* University of Warsaw, Faculty of Philosophy and Sociology. E-mail: ilintar@gmail.com. ORCID: 0000-0003-4714-5269.

representations is that of Dretske, which ties representations with certain natural functions of biological organisms (Dretske, 1986). Dretske's solution has two major strengths. From a philosophical perspective, it can tackle many problems that informational or correlational approaches to representations have had problems with (one of the key issues that the abovementioned approaches faced was that of the ubiquity of representations: smoke is often correlated with fire, but is smoke a cognitive representation for fire? Does fire have cognitive representations?). From a methodological perspective, it can provide a universal take on representations for many classes of organisms, which lets us obtain empirical data about representations from studies on simple animal organisms or even bacteria. It also gives us a clean transition from animal cognition to human cognition—as such, it is a strong counter to all dualistic approaches to cognition.

The main aim of this text is to undermine the universality of cognitive representations as seen by Dretske and his successors. Within this text, we shall be assuming the representational approach in cognitive science. We shall not deal with issues of anti-representationism because, while the question of whether cognitive representations are a valid element of the cognitive science landscape is no doubt interesting and valid, it is out of scope here and would only muddle the main points of the argumentation. Therefore, when we discuss the various pros and cons of the naturalistic approach to representations, remember we do so only under the assumption that representations themselves are useful and significant.

2. Naturalized Representations and Misrepresentation

An important feature that Dretske and many of his successors (e.g. Millikan) emphasize in their solution is the ability to analyze representations with respect to their correctness—in particular, to show criteria of *misrepresentation*. In Dretske's approach, a representation of property X is correct when (in normal conditions) it works according to its function, that being indicating the presence of X. A key aspect of this approach is the concept of function, or more precisely, a specific type of function: a *natural* function. If we want to naturalize representations, we cannot simply use a general notion of function, since there are many possible classes of functions and we risk a problem we seem to have just averted—that of the ubiquity of representations. Therefore, we restrict ourselves to the class of natural functions, which are those that guarantee

evolutionary success. Dretske gives the example of bacteria which have a natural magnetic indicator of north, which allows them to move towards less oxidized waters (a surplus of oxygen is deadly for the bacteria). The same bacteria, when moved to the southern hemisphere, will die, since their magnetic sensor will incorrectly direct them away from the pole, towards deadly waters filled with oxygen.

It's worth noting that even Dretske when providing the example cautions against using it as an instant case of naturalized representation. This is due to the fact that the bacteria's indicator simply points them towards magnetic north, not towards oxygen itself. Even if we assumed an evolutionary criteria for selecting natural functions, it's hard to explain why the mechanism actually indicates the presence of oxygen and not the presence of the magnetic north, with the magnetic north being the environment in which the bacteria normally thrive. In other words, the example with the bacteria moved to the southern hemisphere might not be one of misrepresentation, but one of abnormal world conditions. Even if a given representation could be evoked by one of many independent mechanisms, it still wouldn't be enough to tell us that it's a representation of the property that we desire, rather than a disjunction of the immediate triggers (for example, if the bacteria had a light indicator together with the magnetic one, we could still say that they have a representation of the property Light-or-North and not of the property Oxygen). Dretske claims that only organisms that have a set of independent representation-controlling mechanisms and are able to switch them on during their lifecycle have the capacity to misrepresent. In other words, it is only when an organism has a representation of property X which, during various phases of learning, is evoked by different stimuli originating from X (but having the common feature of being caused by X), that one can talk of a representation that can misrepresent.

It should be clear now that, contrary to the promising start (of bacteria having representations), to talk about representation in the Dretskian sense we need more complicated organisms than bacteria. However, it's still a notion of representation that is scientifically attractive—most animals, even the very simple ones, have some capacity to adapt, so we could obtain a lot of empirical examples for representations and misrepresentation from the rich world of animal behavior.

Moreover, Millikan's solution (1995), which was an answer to Dretske, manages to solve even the problem of bacterial misrepresentation. Millikan solves the problem of vagueness present in Dretske's approach by

assuming that for a given organism, its *proper* function (as Millikan calls her extension of the notion of natural function, see [Millikan, 1987]) is indicating a property that is required for the organism to survive and reproduce (in other words, to achieve evolutionary success). In this approach, the bacteria have the representation of Oxygen (instead of Light-or-North) because it's the former that is required for their survival—the latter is strictly accidental.

All of the approaches mentioned are very well developed and show promise when it comes to studying representations in animals. However, do they actually make it easier for us to understand representation in humans?

3. Human Errors and Humans' Natural Function

Let's now look at a typical case of misrepresentation that happens during the human language acquisition process. A child looks at a ripe, red apple, reaches for it and says "tomato"—with the clear intention of eating the apple as a tomato. She hasn't yet learnt that there are other fruits of similar size, shape and color as the tomatoes that she's observed before.

There are two possible explanations for the situation described. One is that the child simply has a wrong representation of tomatoes, i.e. that she has a representation of tomatoes, but it's not *the correct* representation. Another explanation is that the child does have a representation of tomatoes, but it didn't work correctly that time. Let us call the first explanation that of a *general error* and the latter one—a *particular error*. In both cases we now want to ask the question—how would we naturalize such a notion of representation?

Note that if we want to talk about a functional approach to naturalizing representations (whether it be Dretske's approach or Millikan's approach), we want to talk about a *biological* function—one that we could single out in both humans and in simpler organisms (although, as we mentioned before, Dretske seems to believe that to properly determine a representational function, you need a certain level of biological complexity). This function should be somehow connected with the evolutionary (or, more directly, reproductive) success of the organism. It's worth noting here that Millikan speaks about "representation reproduction" instead of "organism reproduction", which opens up the possibility of understanding it in non-biological terms. However, most of Millikan's own research

pertains to biological reproductive success, so we shall assume that is the dominant understanding for now. We shall tackle the other possibility later in the text.

Now let us consider this: can we actually find a biological function of the child's organism that would determine that the proper representation of the tomato should be one of a tomato and not one of a red apple? Before we actually move on to try to answer this question, it's important to understand that a potentially higher level of complexity (and thus, a more complex function) would not be problematic here. If the difference between human representations and simpler organism representations were just one of degree, that would not be a major difficulty for the naturalized theory of representations. One could rightfully hold the view that the level of complexity of a natural function that realizes a given representation is proportional to the complexity of the organism itself. In such a case we should not find it surprising that a human's natural function is much more complex that one of an amoeba or bacteria. Furthermore, for Dretske such a situation would actually be a pro rather than a con—to talk about natural functions and avoid ambiguity, we need a complex system that makes certain choices based on more than one criterion.

Let us therefore assume that we actually managed to discover a natural function that corresponds to the child's representation of a tomato. Let's also assume that the function actually explains the representational error that the child makes when calling a red apple a tomato (or when it reaches for the apple with the intention of eating it as a tomato, to avoid linguistic criteria). Can any such function really be a *natural* function? Of course, we don't want to define the class of natural functions so widely that it loses its intuitive meaning—after all, we wanted to restrict the class of functions to natural functions precisely to avoid some problems with naturalizing representations. Therefore, we want to relate the natural function to the organism's survival. However, it seems that no credible explanation of that sort can be actually found, as I shall now try to show.

Starting with the most direct approach, a proponent of the naturalistic approach might claim that the ability to distinguish apples from tomatoes is critical for survival. For example, take a child that has a deadly allergy to apples (but not to tomatoes); a misrepresentation might turn out to be fatal (e.g. if the child reaches for the apple and eats it before her parents manage to react). This type of analysis might seem promising, since it only deals with biological criteria. Also, one can provide less convoluted examples where distinguishing one organism from another is critical for

avoiding poisoning. Take, for example, the parasol mushroom and the death cap. This example is even better in that it deals with general mechanisms (the death cap is poisonous for humans as a species rather than just for individuals), so it's easier to claim that such a function would be natural in the sense that it correlates with the evolutionary success of the species.

However, our language is too rich to permit such an analysis for all concepts, so this way is doomed to fail sooner or later. We are not be able to find a direct evolutionary function for every single concept, although we can probably find a scenario in which misrepresenting a concept results in an organism's death. However, inventing scenarios is not a good argumentative road—for every scenario one can find a counter-scenario in which having the allegedly incorrect representation ensures success (for example, a scenario taken almost out of Grimms' fairy tales, where Hansel brings a death cap home and feeds it to the witch, who was just about to cook him in the oven). To justify naturalizing a representation, we must have a universal function—one that can be explained on the level of the entire species, not just single organisms. In the literature, one can indeed find many guidelines on how to correctly describe natural functions so that they are indeed natural (i.e. so that they can be properly naturalized; Millikan's analysis is a good example of this).

We shall drop this line of enquiry now mostly because a criticism of a specific approach to natural functions will not be a definitive rebuttal to the idea of naturalizing representations in general. Even if we cannot tell what the evolutionary advantage is of having the representation of a convertible distinct from the representation of a station wagon, the very fact this distinction exists might suggest that it somehow contributes to our survival. The proponent of the naturalization approach to representations might say that we might not know the exact natural function corresponding to more complex concepts, but it is the task of empirical studies to find and describe it.

Therefore, the objection to naturalizing representation must have a more fundamental nature. The question that will lead us to that objection will be the following: how do we assert misrepresentation in humans? What makes us say that someone misrepresents (in both the general and the particular sense) some class of objects (for example tomatoes or convertibles)? And finally: how do we learn to make the relevant distinctions? The answers to those questions will hopefully cast doubt on the validity of the naturalization approach for human representations.

4. Humans and the Natural Error

Humans are a very specific species in the animal kingdom in that a lot of their representations have a social source—they are created and changed not only in response to stimuli connected to the represented object, but also (or, one could claim, mainly) in the process of socialization. This process of socialization is special even among animals who do have a process of socialization—many of our representations are created with the help of language. I do not want to tackle the topic of the relation between social interactions and cognitive representations in this text, as it would be widely out of scope. This is not only true for representations on the personal level (as per the personal/subpersonal distinction due to [Dennett, 1969]), where the relation to language is quite obvious, but also on the subpersonal level. For example, take the notion of attractiveness—it would seem that the representation of a "potentially attractive mate" is something that we share with the rest of the animal kingdom. However, a short historical enquiry is sufficient to discover that the socially prevalent criteria for attractiveness have changed much more often than would be credible for an evolutionary explanation.

Therefore, even if we restrict ourselves to subpersonal representations, we cannot guarantee that they were not formed without the presence of social factors (unless we are talking about inborn representations; as we shall further discuss, the origin of representations is a quite important differentiating factor). Moreover, if an important feature of representations is supposed to be their durability, then the social explanation seems to be more plausible than the evolutionary one—the example of attractiveness suggests that social interactions are more important in determining representations than purely evolutionary factors.

Let us come back to the definition of misrepresentation formulated earlier and fill in some specific objects for the variables: Athanasius is misrepresenting the parasol mushroom if his representation (parasol mushroom) leads him to collect a death cap in the forest (at least in the general case; in the particular case, he mistakenly takes a death cap to be a parasol mushroom). It would seem that, due to the direct biological effects, this would be a paradigmatic case of naturalized representations—a misrepresentation leads, after all, to an organism's death. However, is this really a case of misrepresentation? More specifically: is the correct representation of the parasol mushroom really what we understand by the linguistic concept "parasol mushroom"? After all, we can

imagine a case where the representation itself does not change, but the inclination to eat the mushroom does. This counter-argument could be rebutted by asserting that having distinct representations for a death cap and a parasol mushroom is evolutionarily superior to having just a representation of a parasol mushroom as an inedible one (again: imagine a scenario in which we have a tribe living in a forest where their only potential food sources are either death caps or parasol mushrooms). However, we can also imagine that the very same tribe represents all those mushrooms as parasol mushrooms—just with the distinction that the greener ones are poisonous, while the more brown-tinted ones are edible. In other words, they ascribe the edibility criteria to certain states of a given type of organism rather than to a distinct type of organism (a real-life case of such a distinction is the mushroom commonly known as a "puffball", whose early forms are actually edible).

Perhaps by now an analogy to a famous argument from philosophy of language—Wittgenstein's criticism of "private language errors" (Wittgenstein, 1953) later expanded upon by Kripke (1982). This analogy does not seem to be accidental—I believe that talking about misrepresentation in the context of our cognitive representations (other than the native ones) in the same way we talk about misrepresentation in the case of simpler organisms in relation to their natural functions is a dead end.

Most arguments that Wittgenstein (and later Kripke) use to refute the possibility of a naturalized conceptual error can be adapted to the case of misrepresentation. Take for example the abovementioned case with death caps and parasol mushrooms. Even Wittgenstein's original example (recall that Wittgenstein, and after him Kripke, claimed that we can't determine whether someone, when talking about addition, or the use of the plus sign, really means "plus" instead of "quus", where quus is different from plus in that it behaves differently in very specific conditions which do not obtain in the given case) could be possibly used (if not for the fact that the concepts used are highly abstract, which makes finding the corresponding representations difficult). Note that the gist of the argument is the same in both cases. Wittgenstein (and Kripke after him) says the following: using purely objective criteria, we are not able to determine, which of the two descriptions of the concept is the correct one—similarly, we cannot determine which of the two descriptions of cognitive representations is the correct one other than rationalizing it *ad hoc* after the fact ("weird parasol mushroom" vs "parasol mushroom / death cap").

This argument can also be used in two ways. If our misrepresentation is understood as a general error (having an incorrect representation), the question becomes: how do we determine the correctness of the representation (in other words, how do we select one specific proper function over all others). If we understand it as a particular error instead, meaning a representation is used incorrectly, then we can ask, after Wittgenstein: how do we know that it was an error and not an exception specified in the rule?

However, the naturalization proponents are in a better place than Wittgenstein's opponents in the rule-based concept usage debate—they can still fall back on the concept of natural functions and defend our representations by relating them to biologically proper functions. However, that route seems a dead end as well—even in the case of the death cap, which seems well-suited for naturalization, it's hard to show a clear advantage of the double representation version over the "weird parasol mushroom" version.

To the fundamental arguments one can add empirical arguments as well. Even if we could agree that the idea of naturalized misrepresentations can be defended on theoretical grounds, it would be hard to defend the claim that our cognitive representations are really formed in the way that this idea describes and that we diagnose misrepresentations based on evolutionary consequences. The richness of our conceptual system and, in consequence, of our representational system (since we have already noted that most of our representations have linguistic correlates) is too big compared to the period of potential evolutionary change for this explanation to actually be plausible. One could defend this type of theory when it comes to bees, whose communication does seem to be evolutionarily coded, but in the case of humans, our systems of communication are too short-lived for the evolutionary context to be relevant.

One could claim that the naturalized approach to representations is nevertheless correct also in humans and that the correct representations are those that realize some natural functions (or proper functions, if we prefer Millikan's terminology) and that the social agreement or disagreement towards concept use has no bearing on the notion of misrepresentation. However, that type of approach requires accepting one of the following assumptions—each of which seems problematic for its own reasons.

First of all, we can assume that linguistic concepts and cognitive representations are not directly correlated—that concepts are not rooted in cognitive representations. In text, we tacitly assume that such a ground-

ing exists, but of course a negation of such a claim can be imagined. In its radical version (concepts have completely no connection whatsoever to cognitive representations) it seems completely implausible for anyone who wants to respect the scientific foundations of cognitive science, including the empirical results of developmental psychology. However, one could opt for a weaker version of the negation—for example, accepting the grounding on the level of types (cognitive representations overall are grounded in cognitive representations), but refusing it on the level of particulars (specific concepts are not grounded in specific cognitive representations). It's hard to see, however, how this type of negation helps alleviate any of the problems mentioned above.

A second option is to assume that the current state of language is not an adequate measurement of the correctness of cognitive representations. Such a solution requires assuming a Leibnitzian view of a perfect language which would best suit our evolutionary needs and which would be the one according to which we should judge representations. However, metaphysical problems notwithstanding, there is a fundamental problem here: is such a solution actually naturalistic? How do we scientifically verify the correctness of representations with a postulated ideal language best suited for our evolutionary success?

The third option is to go holistic—instead of evaluating particular representations as correlated with particular concepts, we evaluate representations based on their role in an entire linguistic system. However, this type of holism only masks the problem—since now we are no longer suited to judge particular representations, instead, we need to evaluate an entire system which the given representation is tied to. This does not seem like a naturalistic approach at all and, moreover, seems to direct us towards an antirepresentational approach which we agreed not to discuss in the introduction.

Besides the problems with the abovementioned three options, the solution that ignores the linguistic side of cognitive representation does not seem to be well reflected in empirical studies. It ignores our actual mechanisms of evaluating representations as erroneous in favor of an idealized concept of misrepresentation that is different from what is commonly understood as a representational error. This seems to be similar to certain solutions within the semantic contextualism debate which, to avoid contextual dependence for some propositions, gives them a literal meaning that comes out as false in virtually all circumstances where we would assert them as true and explains this assertion using a system of complex

implicatures. These types of solutions, while formally correct, do seem dubious in terms of their explanatory power.

5. Representations and Society

Since we have provided the arguments against a fully naturalized solution to representations, a further task remains: to provide an alternative solution to full naturalization. We explicitly refused the antirepresentational solution at the start, so now we are tasked with providing another positive option.

Let us employ a classic tool of analytical philosophy: linguistic use case analysis. When do we say that someone is misrepresenting an object? In our case: when do we say that Athanasius is misrepresenting the parasol mushroom?

We said that, similarly to Wittgenstein's solution, the social consensus seems important here—Athanasius is misrepresenting the parasol mushroom if his representation does not match what society has established as the proper representation. Should we, however, understand this as pure social consent, i.e. Athanasius has the correct representation if and only if society agrees that his representation is correct?

This solution has many benefits and simplifies a lot of matters when it comes to the conceptual side of representations. It's also quite antiscientific—under this approach, we would have to drop all attempts to reduce representations to objects described by empirical sciences. However, that by itself is not a critical problem—after all, we consider many sociological phenomena to be fully emergent, and we don't posit their reduction to the biological layer. This solution has another problem, however—it makes it impossible for us to positively resolve the "individual vs society" dilemma.

Let's consider an archetypal story of a brilliant lone scientist. In this story, an individual comes across a breakthrough discovery, she's shunned by the majority of the scientific community, but then we discover she was right all along. We might try to apply this scenario for example to the real-life historical discussion regarding black holes (assuming that the discussion in this case really was of the "individual vs society" type, since in reality those cases rarely happen in their pure form). If the representation is decided purely by society's consensus, there is no possible case in which the scientist is actually right—she will never be able to prove the correctness of her representation. Even a version of the scenario in which

she gradually convinces the community of her approach isn't well described in this case—since she is constantly wrong when she does the convincing and only starts being right once she's actually convinced the majority. This description seems wrong—we would surely prefer to claim that the scientist was right all along and the majority had the erroneous representation. How can we save this intuitive description?

The best approach seems to be to combine the functionalist approach towards representations, which has a very respectable intellectual history and has developed many useful and precise concepts, with the social approach. To do that, we only have to abandon... naturalization. We would still want to say that having a certain representation is realizing a certain function—what changes is the nature of that function. It would no longer be a natural function—it would instead be a socially-regulated function, in a manner similar to how Wittgenstein understood the way language-meaning rules are governed by society, namely that the meaning of a word is what the linguistic society currently enforces as its meaning.

In such a theory, the scientists who single-handedly maintains the existence of black holes might still be correct—as long as his representational function follows the rules that are enforced by the society. He might still differ with the rest of the society as to what exactly corresponds to the object of those representations (in the same way that I can agree with others that "the fastest man in the world" means "the person who just got the fastest time in the men's 100m sprint at the Olympics", but due to a lack of information I could be convinced that this refers to Justin Gatlin (since that's what the first reports might have indicated), while a later analysis of the photo-finish showed that the fastest one was actually Usain Bolt. This is a bit similar to how Kripke describes necessary truths that are known *a posteriori*—from the fact that the society agrees (explicitly or implicitly) on the meaning postulates regarding a certain concept (in our solution, that would mean they agree with respect to the representation function that realizes the concept), it does not follow that they have knowledge about all true propositions which the concept is part of, as some of those propositions can be only known by empirical research and not just by conceptual analysis. Our token scientist might therefore agree with other scientists on the ostensive definition of black holes (e.g. "black hole" = "that which constitutes the center of known galaxies"), while disagreeing on the essential physical properties of those objects (e.g. their ability to capture light or alter the gravitational field).

If someone still remains unconvinced by the analysis above, here's an alternative argument showing that the naturalistic approach to human representation is not plausible—one grounded in the results of cognitive sciences (some elements of that argumentation can already be found above). Let us consider what is the subject of inquiry of cognitive science when it comes to humans and compare it to the subject of inquiry when it comes to animals and other living organisms. Assuming that we can provide a common metaphysical description of representations in both cases, we have to ask how the respective representations are formed. It seems that while in the case of animals almost all representations are inbred and have an evolutionary source, that's not the case with humans—our representations, judging e.g. from their linguistic correlates, seem to be contingent and have a social ground. If we aim at providing a common characteristic for human and non-human representations by using natural functions, we obtain easy empirical data, but the data will not necessarily be adequate for human cases (since it's hard to provide a credible evolutionary explanation for humans in the same way that it's possible for simpler organisms). Therefore, this solution picks data accessibility from the accessibility / credibility pair, which of course is better if we want an easy influx of superficially convincing examples, but raises concerns from a methodological perspective.

Let us come back to the solution offered by Millikan that we mentioned earlier and see if we can recover the naturalistic approach by assuming a broader approach to the concept of representation reproduction. Can we understand "reproduction" in a social way here and assume that representations are persistent if they are socially reproducible? Of course, we could do that, but it seems that for a naturalist that would actually be a pyrrhic victory. While it seems quite obvious that biology is the science that is suitable for describing evolutionarily stable mechanisms, it would be quite a stretch to assume that biology is likewise suitable for describing socially stable mechanisms (such as linguistic concepts, language systems or cultural norms). We can refer to sociology, psychology or economics to fuel us with theories that handle those concepts, however, it will be hard to assume that such a solution will still be naturalistic. Usually, by "naturalistic" we understand a reduction to the results of natural science—we would either have to use an unusually broad notion of naturalization or assume that sociology, psychology or economics can be completely naturalized—which would be defending the naturalization

of representations by assuming an even stronger and more controversial claim.

Therefore, if we want to defend the functionalist approach to representations, we shall have to modify many assumptions that usually underlie this approach. Most importantly, we shall have to get rid of the "natural" teleology and the corresponding approach to natural functions which ties them with evolutionary stability (an approach common in Millikan's writing). An in-depth analysis of the argumentation provided in this paper allows us to go even further—we should get rid of teleology completely. An analysis that takes into account both representation data from human and non-human examples suggests that we might be better off by instead considering proper functions with respect to their causes instead of their purposes. This would also help explain the teleological approach present in the research on representations—in the evolutionary approach, cause and purpose are almost indistinguishable (it is very hard to tell "has function X because his genes survived" apart from "has function X to allow his genes to survive"). However, the two categories are very sharply distinct when it comes to human representations—we can talk more easily about representations that have a linguistic origin and distinguish them from ones which have an evolutionary origin (note that under such an approach, we do not assume that there are no evolutionarily-driven representations in humans—again, taking into account the results of cognitive science, such an assumption would be quite controversial, as many subpersonal representations, especially of the simple perceptual variety, do seem to have an evolutionary origin).

6. Summary

In light of all the argumentation presented, it seems that a completely naturalistic approach to human representations is hard to defend. Not only are there good philosophical reasons to refute it, there are also strong methodological reasons for the refutation related to the origin of representations in humans. On the other hand, the hybrid social-functional approach sketched here, which uses the origins of representational functions instead of their purposes, seems to be better suited for explaining the differences in representations between humans and simpler non-human organisms, as well as for dealing with the problem of misrepresentation. It remains to be seen how much of the research on functionalism with respect to representations can be ported to such an ap-

proach—however, I believe that such a hybrid solution would be effective and have the added benefit of bridging the gap between naturalistic and anti-naturalistic approaches to cognitive representations.

REFERENCES

Dennett, D. (1969). *Content and Consciousness*. London: Routledge.

Dretske, F. I. (1986). Misrepresentation. In: R. Bogdan (Ed.), *Belief: Form, Content, and Function* (pp. 17–37). Oxford: Clarendon Press.

Kripke, S. (1982). *Wittgenstein on Rules and Private Language*. Cambridge, Mass.: Harvard University Press.

Millikan, R. G. (1989). Biosemantics. *The Journal of Philosophy*, *86*(6), 281–297.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, Mass.: MIT Press.

Millikan, R. G. (1995). Pushmi-Pullyu Representations. *Philosophical Perspetives*, *9*, 185–200.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell Publishing.

Originally published as "Reprezentacje znaturalizowane – użyteczny cel czy użyteczna fikcja". *Studia Semiotyczne*, *31*(1), 91–108, DOI: 10.26333/sts.xxxi1.06. Translated by Piotr Wilkin.

Studia Semiotyczne—English Supplement, t. XXX (2019), s. 21–41
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/stsen.xxx.02
Article

Jacek Wawer *

THE PROBLEM OF INDEX-INITIALISATION IN THE TEMPO-MODAL SEMANTICS

Summary: In Kripke-semantics for modal logic, the truth value of a sentence depends on the choice of a semantic index (e.g. world, time, or place). It means that application of such semantics to natural language analysis requires indication of an index relevant for semantic analysis. It is commonly accepted that the relevant index is initialised by the context of an utterance. The idea has been rejected by the semanticists investigating tempo-modal languages in the framework of indeterminism, which generated the problem of initialization of the semantic index. I present the main argument of those semanticists and describe several responses to the initialisation problem. I finally argue that under certain metaphysical and semantic assumptions, one can respond to the initialisation failure in the classical way, even in indeterministic contexts.

Keywords: future contingents, semantics of modal languages, context dependence, modal metaphysics.

The truth value of the sentence "It is snowing in Cracow" depends on the time. The truth value of the sentence "There are mountains around" depends on the place. The truth value of the sentence "Pigs fly" depends on what the world is like. When modelling this phenomenon using Kripke's semantics, we postulate that the semantic value of expressions can

* Jagiellonian University, Faculty of Philosophy. E-mail: jacek.wawer@uj.edu.pl. ORCID: 0000-0003-2546-0962.

change along with the semantic index. In the semantics of temporal operators, sentences can have different values at different moments. In the semantics of spatial operators, sentences change their truth-value depending on the choice of spatial coordinates. In the semantics of possibility and necessity, sentences can assume different truth-values in different possible worlds. When we work with multimodal language, the semantic index must be rich enough and contain a parameter for the interpretation of each modality: a parameter of world for necessity, a parameter of time for temporal modalities, a parameter of place for spatial modalities etc.[1]

The variability of semantic value of an expression along with a changing semantic index parameter is an essential feature of semantics for modal languages as the function of modal operators is nothing else but shifting an appropriate semantic index parameter. For instance, the operator of possibility changes the parameter of possible world: in a world w, the sentence "Pigs could fly" is true if and only if the sentence "Pigs fly" is true in a world w', accessible from the world w. Similarly, temporal operators change the parameter of time: The sentence "It was snowing" is true at the moment t if and only if the sentence "It is snowing" is true the moment t', which is earlier than the moment t. One can say that within the semantics for modal languages respective parameters of the semantic index must be "mobile".

The classic semantics of quantifier logic has a similar feature; in this case, the changeable parameter is the valuation function. Just as in modal semantics the semantic value of a sentence can change along with a change in the world, in the semantics for quantifier logic the semantic value of the formula $P(x)$ can change along with the changes to the valuation function (a formula can be satisfied by one valuation and not fulfilled with another). The analogy reaches even deeper; notably, as the main function of modal operators is to shift the modal parameter of the semantic index, the main function of quantifiers is to shift (i.e. appropriately change) the valuation function.

[1] In the entire text, I will interpret modal modifiers as sentence operators rather than quantifiers, even though this assumption is disputable (see e.g. King 2003). I assume this rather for the simplicity of exposition than out of deep conviction. I need to stress, however, that the problem of index initialisation discussed in the text arises regardless of the choice of the formal representation of modality.

This is, of course, not a full analogy. On the formal level, it is easy to notice that the valuation function is a much more subtle tool than the possible world. For example, it allows for independent quantification over different variables while the modal operator has only one possible world "at its disposal". The analogy between the semantics of modal operators and the semantics of quantifiers also breaks at the level of application to natural language analysis. Within the language of quantifier logic, there is a common distinction between open formulas and sentences (closed formulas). The difference is that in an open formula there is least one free variable (beyond reach of any quantifier). Most, if not all, sentences of natural language that can be translated to the language of quantifier logic become closed formulas after translation (except for, maybe, sentences like "This is white" where the context does not specify what exactly is meant by "this").

The fact that the typical natural language sentences translate to closed formulas is consequential when we apply logic to the analysis of natural language sentences. An important feature of the quantifier language semantics is that the truth-value of the closed formulas, contrary to the open formulas, is independent of the valuation function. This means that while an open formula can change its semantic value depending on the valuation function (it can be satisfied or not), a closed formula is satisfied with every valuation if it is satisfied with one (and if it is not satisfied with one valuation, then it is not satisfied by any other). The valuation function parameter is the key supporting tool, which makes the recursive definition of the satisfaction function possible, but on the level of assigning semantic values to closed formulas (i. e. sentences), its value ceases to be relevant. Thanks to this characteristic of closed formulas, semantic analysis of the sentences of natural language using the tools of quantifier logic is uncontroversial. Even though, for the sake of uniformity, the valuation function should be indicated to assess the semantic value of a sentence of language, we are not forced to specify which particular valuation function is "right" as the choice of one or another function is irrelevant.

The situation changes drastically if one tries to use modal logic to analyse sentences of natural language. Elementary formulas of the language of modal language represent sentences like "It is raining", "Pigs fly" and their truth-value *depends* on the choice of an appropriate semantic index parameter (world, time, place etc.). Hence, while the choice of valuation function is not relevant to the semantic values of the sentences of quanti-

fier logic, the choice of the possible world, place or time has key influence on the semantic value of sentences like "It is raining".

Thus, anyone intending to use the formal apparatus of modal logic for semantic analysis of sentences of natural language is confronted with the question: which of the modal parameter values should be chosen to assign semantic values to the sentences of ordinary language? I am going to call this question the problem of index initialisation.

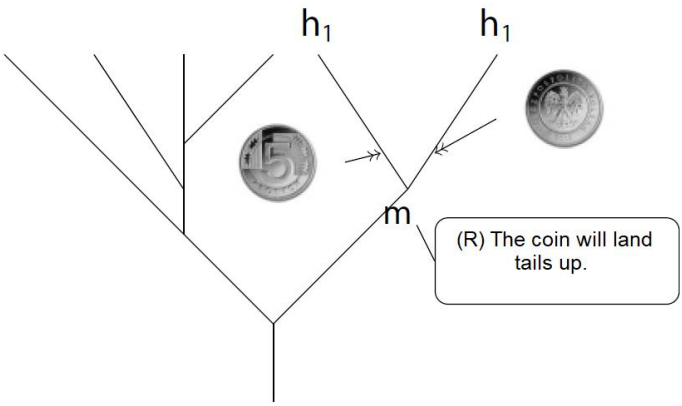
An answer to a problem so stated was outlined already by Kazimierz Twardowski, who addresses a similar issue in his paper *On the So-Called Relative Truths*: "Circumstances accompanying the utterer's words supplement what the words do not express" (Twardowski, 1900, p. 68; translation by Agnieszka Przybyła-Wilkin). In the contemporary literature, the "circumstances accompanying the utterer's words" are usually called context and the "supplementing" Twardowski writes about will be called "index initialisation" by me. Twardowski presents a very natural solution to the problem formulated above: if the truth-value of a sentence depends on the choice of the semantic index parameter, then this parameter is initialised by the context in which the sentence is uttered. Thus, to assign the truth-value to the sentence "It is raining" uttered on top of the Castle Hill in Lvov on the 1st of March 1900 (in our world), one should choose the following parameters: the Castle Hill in Lvov as place, the 1st of March 1900 as time, and our world as possible world. This approach was popularised by David Kaplan, who, in the commentary to his groundbreaking work *Demonstratives*, strongly emphasised the double role of context: as a tool to interpret occasional expressions and as a tool to initialise the appropriate semantic index for interpretation of natural language sentences. (see Kaplan, 1989, p. 595). It seems that, thanks to the support of the context of an utterance, the problem of index initialisation disappeared as quickly as it had appeared. However, not all philosophers are fully satisfied by this answer.

Problem with the World of Context

The answer by Twardowski-Kaplan to the problem of index initialisation has been questioned in the context of semantics created to analyse time-dependent possibilities (possibilities that vanish with time). A good tool to examine these possibilities turned out to be the model of branching histories (or worlds). This model assumes that histories can overlap in an initial interval and then part ways. The history that "branched" in the

past of a given point represents the possibility that was accessible in the past but vanished as the time passed. Such possibility can be exemplified with a history in which the citizens of Great Britain vote for remaining in the EU. It was available before the referendum, which took place on the 23rd of June 2016, but the real development of the referendum annihilated that possibility. Mutual relations between histories can be pictured as a tree, as Figure 1 shows (the first version of the model of branching histories was proposed by Arthur Prior, who was inspired by Saul Kripke's suggestions, see Prior, 1966; 1967; Øhrstrøm, 2012).

Figure 1



The language in which we want to talk about temporal modalities contains temporal operators: "it will be the case that", "it was the case that" as well as the operator of historical necessity: "it is inevitable that". To interpret these modalities, we need two parameters in the semantic index: the parameter of time and the parameter of history (or world). While the initialisation of the time parameter by context does not raise any serious doubt, the initialisation of the history parameter turned out to be much more controversial. Consider a sentence (R) "the coin will land tails up" uttered at the moment m indicated in Figure 1. While it is clear which time is initialised by the context—the time at which the sentence (R) is uttered—it is not clear which history (world) should be initialised. Our answer cannot be analogous—the history in which the sentence (R) has been uttered—because the sentence (R) has been uttered both in the history h_1 and in the history h_2. This fact was emphasised by several authors: "Unlike worlds, histories overlap, so that a single speech

act will typically belong to many possible histories" (Belnap, Perloff, & Xu, 2001, p. 152), "the utterance takes place in many worlds" (MacFarlane, 2008, p. 85). Consider a concrete case in which a sentence is used [...]. There will be many worlds, in general, that represent the very same past and present happenings [...]. The concrete episode of use takes place in all of them." (MacFarlane, 2014, p. 208).

However, if the sentence is uttered simultaneously in a number of different histories, it is not possible to indicate "one history in which the sentence has been uttered". As a consequence, there is no simple method to indicate "the only history of the context", which restores the problem of index initialisation. Semantics requires an indication of an index—of time and of history—to commence the analysis of the sentence "The coin will land tails up", while metaphysics does not allow us to distinguish any index.

We cannot use the argument that helped us with quantifier semantics. Semantics for quantifier logic also requires indication of a certain valuation function to allow semantic analysis of a sentence. Obviously, no such function is determined by the context of the utterance. In the case of quantifier logic, however, it quickly turned out that it does not matter which function we indicate as the semantic value of a sentence (closed formula) is independent on the choice of valuation function. This is not the case here. The semantic value of the sentence "The coin will land tails up" is dependent on the choice of the history parameter. This sentence is true in the history h_1 but false in the history h_2. however, this value is not established by the context. Thus, it turns out that the application of modal logic semantics for analysis of natural language sentences brings about a fundamental difficulty, particularly if we focus—as in our example—on future contingents.

(Post)Semantics of the Future

To tackle this problem, philosophers and logicians suggested a wide array of solutions. The first attempt was made by Arthur Prior who defined the semantics he called Peircean (Prior, 1967).[2] The Peircean theory gives up the operators of possibility and necessity while it enriches the temporal operators with the component of necessity. In ordinary temporal

[2] The name is a reference to the thought of Charles Sanders Peirce, whose writings inspired Prior's solution.

logic, we shall say that the sentence "The coin will land tails up" is true at the moment m if and only if the sentence "The coin is landing tails up" is true at the moment m' later then m. Peircean semantics modifies this condition, saying that:

> The sentence "The coin will land tails up" is true at the moment m iff in every history the moment m belongs to the sentence "The coin is landing tails up" is true in a moment m' later than m. Otherwise, it is false.

Thus, the difficulty with indicating the right history is solved by quantifying over all histories, which results in the operator "it will be the case that" containing a component of necessity "it is inevitable that it will be the case that". Such an alteration of meaning, however, makes the Peircean semantics a worse tool for the analysis of grammatical tenses. For instance, in Peircean semantics, before the coin toss, the sentence "The coin will land heads up or tails up but it will land neither heads up nor will it land tails up" ($F(p \vee q) \wedge \neg Fp \wedge \neg Fq$) is true, even though it sounds like a contradiction. To see that, one just needs to look at the model depicted in Figure 1. The sentence $F(p \vee q)$ is true at the moment m because in each history going through m, there is a later moment in which it is true that the coin lands heads up or tails up ($p \vee q$). At the same time, both sentences Fp and Fq are false at the moment m because the coin does not land tails up in all histories and does not land heads up in all of them.

Another suggested solution to the index initialisation problem is to adapt Łukasiewicz's trivalent logic to the models of branching histories.[3] In this adaptation we will say that:

> The sentence "The coin will land tails up" is true at the moment m iff in each history the moment m belongs to, the sentence "The coin is landing tails up" is true at a moment m' later than m.
>
> The sentence „The coin will land tails up" is false at the moment m iff in each history the moment m belongs to, the sentence "The coin is landing tails up" is false at every moment m' later than m.

[3] Interestingly, Prior introduced his tense logics—Peircean and Ockhamist—as an answer to Łukasiewicz's trivalent logic, which he had earlier defended. However, Prior's logics were hard to accept for Łukasiewicz because the logical operators present in them are extensional.

> Otherwise, the sentence "The coin will land tails up" assumes the third truth-value.[4]

The fundamental difficulty of the trivalent semantics, however, is the fact that the sentence "The coin will land tails up or it will not land tails up" is assigned the third truth-value while we intuitively deem it true.[5]

An innovation allowing us to solve this problem is Richmond Thomason's (1970; 1984) semantics of supervaluations. In this solution, Thomason employs two kinds of valuations simultaneously. Basic bivalent valuations assign classic truth-values to sentences in relation to the moment/history pairs and supervaluations assign truth-values to sentences in relation to moments only, according to the pattern explained below. The supervaluation technique allows for introduction of (super)truth-value holes while keeping the tautologies of classical, modal, and temporal logic. Thomason's solution was inspired by the work of Bas van Fraassen (1966).

Analysing Thomason's solution, John MacFarlane (2003) identifies the (super)truth at a moment with truth at the context and the truth in the pair moment/history with truth at the index. He argues that the latter notion has only a supporting function. Its purpose is to clarify the earlier one, which should model our immediate intuitions concerning the truth-value of sentences uttered under concrete circumstances. MacFarlane calls the theory connecting the notion of truth at the context with the notion of truth at index "postsemantics". In this terminology, postsemantics of supervaluations addresses the problem of index initialisation as follows:

> The sentence "The coin will land tails up" is true at the context m iff it is true at every index $\langle m, h\rangle$ where m is an element of h.
>
> The sentence "The coin will land tails up" is false at the context m iff the sentence "The coin will not land tails up" is true at the context m.
>
> Otherwise, the sentence "The coin will land tails up" lacks the truth-value at the context m.

[4] This is a definition proposed by John MacFarlane (2014, ch. 9.6); an alternative definition, preserving the extensionality of all operators, can be found in (Wawer, 2016, ch. 4.4).

[5] Notably, Łukasiewicz himself agrees in his *On determinism* that we should assess this sentence as true (see Łukasiewicz, 1961, p. 124).

If we apply this definition to the example in Figure 1, it turns out that at the context m the sentence "The coin will land tails up" is neither true nor false. At the same time, the sentence "The coin will or will not land tails up" is true at the context m (as it is true in every history running through m).

A problematic consequence of the semantics of supervaluations is the fact that the classic logical connectors (like disjunction) are not extensional at the context. In the above example, an alternative of two sentences without truth-value is true but we can easily find examples where an alternative of two such sentences has no truth-value. For instance, if I make a wager that the coin will land tails up, the disjunction "The coin will land tails up or I will win the wager" has no truth-value.

Another problem of the semantics of supervaluations, particularly stressed by MacFarlane, is that although the sentence "The coin will land tails up" has no truth-value at the context m, at the later context m', which belongs to the history h_1, the sentence "It was true that the coin would land tails up" is true. MacFarlane states that this characteristic leads to counterintuitive consequences. MacFarlane's objection is very subtle and has changed its form over time (see MacFarlane, 2003; 2008; 2014). Therefore, I will not delve into details here. A summary of the discussion can be found in Wawer 2016, ch. 4.6.

MacFarlane's answer to the problems of postsemantics of supervaluations is his own assessment relativism. According to this theory, the semantic value of a sentence should be established upon consideration of not only the context of the utterance, but also the context of assessment.

Coming back to our example, the sentence "The coin will land tails up" uttered at the moment m has no truth-value when assessed in the context m; when assessed in a later context within history h_1, it is true; when, in turn, assessed in a later context of the history h_2, it is false. This effect is achieved by MacFarlane thanks to the following definition:

> The sentence "The coin will land tails up" is true at the context of utterance m and the context of assessment m' iff it is true at every index $\langle m, h\rangle$ at which m' is an element of h.
>
> The sentence "The coin will land tails up" is false at the context of utterance m and the context of assessment m' iff the sentence "The coin will not land tails up" is true at this pair of contexts.

> Otherwise, the sentence "The coin will land tails up" has no truth-value at the context of utterance m and context of assessment m'.[6]

MacFarlane makes a case for his semantics by referring to our intuitions on accuracy of utterances. He argues that the act of uttering "The coin will land tails up" is not accurate before the coin toss, while after the toss in which the coin has landed tails up, that very same act of uttering is accurate (or, more precisely, was accurate). This can be explained by indicating that the sentence uttered before the toss is not true in the earlier context of assessment but is true in the later one (assuming that the truth is a necessary condition of the utterance's accuracy, i.e. truth is a norm of assertion). I have a number of doubts concerning both the diagnosis and the treatment proposed by MacFarlane. Commenting on my doubts, however, would take us too far away from our main point; I will therefore leave this comment for another occasion and move on to one more reaction to the problem of index initialisation.

This reaction is presented by Belnap, Perloff and Xu (2001, ch. 6C). According to them, asking for the semantic value of the expression "The coin will land heads up" at the context m is simply nonsense. They compare the expression "The coin will land heads up" to the formula "x is white". Just as in the latter case there is no sense in asking whether the formula is fulfilled without indicating a valuation, it makes no sense in the earlier one to ask about the truth of the expression without indicating a suitable parameter of history. On the other hand, when we do indicate the suitable parameter, the answer is simple: "x is white" is true with respect to a valuation that ascribes snow to "x" and "The coin will land heads up" is true when we choose a history in which the coin lands heads up as a parameter of evaluation. Thus, we can think of the expression "The coin will land heads up" as a formula containing a free variable ranging over the set of histories. One can assume that the deep structure of this expression actually has a form "In the history h the coin will land heads up", where h is a variable.

What causes my uncertainty towards such an analysis is the fact that we do not usually think of the expression "The coin will land heads up" as a sentence function, which changes its value depending on the arbitrarily chosen value of the parameter h. We rather consider this expression

[6] If there are no histories containing both m and m', the truth-value at the pair of contexts is reduced to supertruth at the context m.

a full-fledged sentence, which, after specifying the moment of utterance, is truth-apt. In everyday practice, we do not even get the idea that the sentences we utter about the future cannot be ascribed a truth-value unless one of the possible future histories has been indicated (not to mention that it is not quite clear what the indication of a possible history should look like).

Moreover, while no sensible person will use the expression "x is white" to communicate a thought, we do not have problems using sentences like "The coin will land heads up" or "Next week I will be in Lublin". One of the explanations of the lack of analogy is that (contrary to Belnap's argument) in the first case, one cannot sensibly ask for the truth-value of these expressions, while in the other two one can do it. Belnap, Perloff and Xu (2001) propose an alternative explanation to this discrepancy (this answer is discussed in more length in [Belnap, 2002]). They believe that the difference on the pragmatic level—we assert sentences about future, we do not assert open formulas—stems from a different *modal profile* of the two cases. Even though the formula "The coin will land tails up" is neither true nor false, it will *have been decided* in the future that the sentence was true or it will *have been decided* that it was false.[7] One can say that over time, a sentence uttered today becomes independent of the choice of history parameter, which makes it usable in the language practice. However, instead of a detailed description of Belnap's ideas, I will suggest an alternative answer to the index initialisation problem.

The Possibility of Futurism

Contrary to the well-established opinion among the researchers of branching histories, I will argue that one needs not reject the natural interpretation of temporal operators or change logic to answer the problem of index initialisation. I believe there is no reason not to refer to the context as a source responsible for establishing both the time and the history, even considering the undetermined future. I will opt for Twardowski-Kaplan's conservative answer to the index initialisation problem.

I think that the impression that the model of branching realities precludes the traditional solution to the problem of index initialisation stems

[7] This observation of Belnap's inspired MacFarlane to create the assessment relativism.

from a very special interpretation of this model, which I call "branching realism". According to this interpretation, alternative histories in some ways resemble David Lewis's possible worlds (see Lewis, 1986). Like Lewis's worlds, all histories are equally real and metaphysically on par with the history (or histories) we belong to. All histories consist of concrete events and none of them is metaphysically distinguished. Moreover, just as the inhabitants of each of Lewis's worlds can rightfully say about their world that it is the actual world, the inhabitants of every point in the tree can rightfully say that their situation is actual.

Although the theorists of the branching model try to avoid unequivocal metaphysical declarations, many of them suggest that their reflection is based upon such realism. One of the branching theory classic authors, Richmond Thomason, writes:

> Consider two different branches, b_1 and b_2, through t, with $t < t_1 \in b_1$ and $t < t_2 \in b_2$. From the standpoint of t_1, b_1 is actual (at least, up to t_1). From the standpoint of t_2, b_2 is actual (at least, up to t_2). And *neither standpoint is correct in any absolute sense.* (Thomason, 1984, p. 145, emphasis added)

Then he adds:

> See D. Lewis (1970), and substitute "the actual future" for "the actual world" in what he says. *That* is the view of the thorough-going indeterminist. (Thomason, 1984, p. 145, note 14, emphasis in original)

Belnap, Perloff and Xu write in a like spirit:

> To suppose that there is one from among the histories in *Our World* [as the authors call the branching model—J. W.] that is the absolutely actual history is rather like purporting to stand outside Lewis's realm of concrete possibilia and pointing to the one that is actual. But this is wrong in both cases. (Belnap, et al., 2001, p. 163)[8]

Some statements by John MacFarlane also suggest modal realism:

> There is nothing in the branching model that corresponds to a car moving along the branching road, and nothing that corresponds to the decision the

[8] There is also a realistic overtone to their definition of "Our World", which can be found in Belnap, et al., 2001, pp. 139–140.

> car will have to make to go down one branch or the other. If worlds branch, then *we branch too*. (MacFarlane, 2014, p. 212, emphasis in original)

A similar metaphysical vision transpires from the semantic objections by Belnap and MacFarlane cited above. The authors agree that a concrete utterance is a part of many different histories/worlds. Such vision is also suggested by Figure 1. The image shows the utterer as an inhabitant of a tree whose all parts are as real as the speaker and their utterance.

It is worth noting here that the realism of branches is in some significant ways different from David Lewis's realism of worlds. First of all, histories (lines in a tree) overlap while Lewis's worlds are disjoint. It is, however, noteworthy that Lewis's attitude to overlapping worlds is not unequivocally critical. He believes that the worlds so understood are in opposition to some common-sense statements (Lewis, 1986, pp. 207–208; incidentally, I believe that Lewis is wrong in this respect). However, he also notices that realism so understood can relatively easily solve problems with which he himself had to struggle (such as the problem of transworld identity, see Lewis, 1986, p. 209). He also adds that

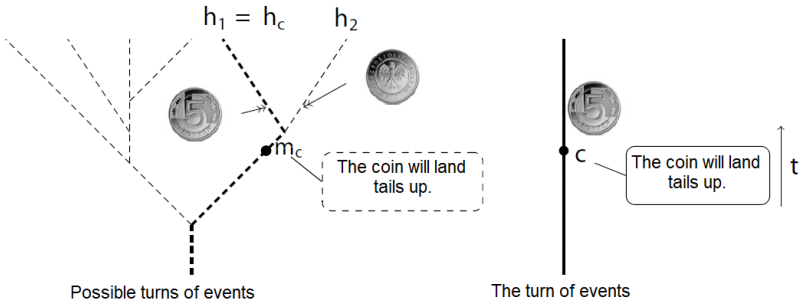
> Overlap spoils the easiest account of how worlds are unified by interrelation: namely, the mereological analogue of the definition of equivalence classes. But alternative accounts are available [...], so I presume that a modal realist who wished to accept overlap would not be in serious difficulty on this score. (Lewis, 1986, p. 209)

The realists advocating branches also distance themselves form some of Lewis's views (see esp. Belnap, et al., 2001, ch. 7A.6) but in general, they have more similarities than differences. In particular, they agree that an absolute distinction between the actual and the possible is wrong. They believe that each possibility is actual from its own perspective and none of the modal perspectives are distinguished.

This is, however, not the only available interpretation of the structure of branching possibilities. Instead of accepting Lewis's vision of possibility, one can join Adams (1974), Plantinga (1976) or Kripke (1980) and accept some form of modal actualism. From the point of view of our problem, the key aspect of modal actualism is the postulate of *absolute*, i. e. not only relative, distinction between the actual and the possible. Contrary to what Lewis states, possible worlds are not metaphysically similar to the world we live in. The world which we belong to is an entity of a different nature—an entity that *realizes* one of the possibilities. With such an ap-

proach, the branching model is a visualisation of *possible* temporal evolutions of our world, yet these possible evolutions are fundamentally different from the world we are part of. Importantly, while possibilities branch over time, the world does not have a branching structure like this. It evolves in a linear manner and with time, it fulfils one of the possibilities available.

Figure 2



When we adopt such perspective, the problem of index initialisation is seen in a completely different light. We cannot say, like the modal realists, that a specific act of uttering is a part of many possible histories. Utterances do not occur in possibilities but in the concrete reality. Every such utterance is a part of only one world ("our" world). This situation is visualised by Figure 2. The line on the right shows all the events that have occurred, are occurring and will occur in our world while the tree on the left shows all possible courses of events. One of the possible courses of events is, of course, the way events have actually turned out and will turn out (the world evolves in a "consistent" manner, realizing only one of the possibilities). I marked this possibility with a bold line.

Such take on the relation between the actual and the possible allows for a completely different answer to the problem of index initialisation. Contrary to what Belnap and MacFarlane say, the utterance of the sentence "The coin will land tails up" does not take place in many different histories/worlds. The utterance occurs in exactly one world, which allows us to return to the standard answer to the index initialisation problem: the world of context is the world in which the utterance takes place. More precisely, the world/history indicated by the context is the only possible history accurately representing the way the world was, is and will be.

Since the world evolves in exactly one of the possible ways, it is guaranteed that there is only one history that accurately represents this evolution. This is the history that should be chosen as the history of context. On the intuitive level this comes down to the trivial observation that the sentence "The coin will land tails up" is true if and only if the coin actually will land tails up in the future, which can be more formally presented in a form of a statement I call "futurism":

> The sentence "The coin will land tails up" is true in the context c iff it is true at the moment of context m_c and at the history of context h_c.

The moment of context is by default the present moment and the history of context is by default the actual history. One can, therefore, answer to the problem of index initialisation in the conservative style of Twardowski, even if the sentence analysed is a future contingent. However, in order to do this, we need to refer to the metaphysical principle of actualism: that the world which we belong to (and in which we utter sentences) is metaphysically of a different nature from the *ways* the world can evolve. When adopting such assumption, we can defend our argument against the objections of modal realists, raised against the notion of the world of context.

I achieved the connection between the metaphysics of actualism and semantics through observation that acts of uttering are a part of one specific world, which differs in nature from the possible evolutions. One can, however, object to this statement as follows: even the actualists, who distinguish metaphysically between actuality and potentiality, often accept a paraphrase of modal sentences in categories of possible worlds. Moreover, they will not have a problem accepting the statement that some utterances that never have taken place and never will take place, could have taken place. For instance, Senator Elizabeth Warren could have backed Bernie Sanders in the 2016 Democratic Party Presidential Primaries, yet she did not. Thus, even actualists are eager to admit that there is a possible world in which Elizabeth Warren utters the sentence "I shall do everything for Bernie Sanders to become the next president of the USA." Therefore, contrary to what I stated above, even within actualism, utterances are present not only in our world, but also in the possible worlds. If this, in turn, is true, our world has not the exclusive right to

utterances and so it cannot be used to solve the semantic index initialisation problem.[9]

One can answer a difficulty put that way in one of two manners: elitist or egalitarian. In the earlier strategy, we focus on the special status of our world and negate the statement that any utterances occur in any other possible worlds (this is the strategy I suggest in [Wawer, 2014]). The statement that there is a world in which Senator Warren says anything is, after all, just a useful *paraphrase*, or *metaphor*. What is paraphrased depends on the specific version of modal actualism. The statement that there is a possible world in which Elizabeth Warren says "A" could be, to name a few examples, be understood as follows:

- Elizabeth Warren could have said "A".
- E.W. had a disposition to say "A".
- There is an (abstract) non-contradictory set of propositions that represents E. W. saying "A".
- There is an (abstract) maximal state of affairs, part of which is E. W. saying "A".
- There is an (abstract) way the world could have been within which E. W. says "A".

What is important to us is that *none* of these paraphrases suggest that besides specific acts of utterance, which take place in our world, there are similar acts occurring in other worlds. For instance, the proposition that E. W. utters the sentence "A" is an entity radically different in its nature from a real utterance of the real E. W. Therefore, we need not be troubled with the acts of utterance taking place in other worlds as, literally speaking, such acts do not exist (there are only states of affairs or propositions representing such acts, dispositions to such acts, possibilities of such acts occurring etc.). Our task was to indicate a mechanism that connects a specific utterance with a suitable semantic index relevant for the semantic interpretation of this utterance. Since utterances take place only in one world, we have a full guarantee that the context of the utterance will establish the appropriate semantic index (actual history and present time). An elitist actualist of this type must, of course, face the obvious observation that E.W. could have said "Bernie Sanders will be

[9] I thank an anonymous reviewer for raising this objection.

the next president", or even that E. W. could have truly said "Bernie Sanders will be the next president". However, the analysis of the sentence "E. W. could have truly said 'Bernie Sanders will be the next president'" does not require us to assume that in some other place, E. W. really utters the sentence "Bernie Sanders will be the next president" (a proposed analysis of reports of utterances embedded within reach of modal operators can be found in [Wawer, 2016, ch. 6.3.6]).

One can also propose a more egalitarian, conciliatory line of answer to the difficulty outlined above. In this strategy, we approach the possible utterances more sympathetically and agree that every such utterance can be *treated as if* it was factual—or, more precisely, only the factual utterances take place but one can assume, or imagine, that a given utterance is factual and formulate a problem analogous to our index initialisation problem: Assuming that Elizabeth Warren indeed says "Bernie Sanders will be the next president", which of the histories running through this possible utterance should be used for the semantic analysis of her utterance? The problem might seem very acute as I have argued earlier that it is the particular, factual world and its turn of events that establishes the possible history relevant for the process of semantic analysis and in our example I explicitly assume that E. W.'s utterance is not a part of this world (E. W. never actually uttered these words). Thus, the possible situation of utterance "lacks a world" that could help us establish the appropriate semantic index.

I believe that a difficulty of this type stems from a misunderstanding whose root is a kind of doublethink: on one hand, we treat the utterance of E. W. as if it was factual while on the other hand, we stress that it is merely possible. This kind of vision is indeed problematic and leads to controversial conclusions.[10] Nevertheless, an actualist need not, or even should not adopt it. If we prefer the egalitarian approach to the branching model, we decide to assume that every possible situation can be the context of utterance. Still, in this case we should remember that when treating a given possible situation as the context, we must also assume that this situation is actual and, as such, it is a part of the actual course of events, which realizes one of the temporal possibilities available at the moment of utterance. If it is so, then the semantic index can be initialised

[10] Notably, this very kind of doublethink is spread among the critics of actualism in the context of the branching model, such as Nuel Belnap or John MacFarlane.

in the exact same way as we initialise it in the case of actual utterances. The time of context is the present time of the utterance and the history of context is the history that will be satisfied by the course of events containing the considered utterance. Hence, whether we adopt the elitist or the egalitarian attitude to the possible utterances, we reach a conclusion that when analysing semantically the utterance used in the given context, we must treat it as a part of the actual world and therefore, we can refer to this world to establish the appropriate semantic index.

It is worth noting that accepting Kaplan's traditional solution to the index initialisation problem, we take one side of the conflict going back to the ancient times about the truth-value of the future contingents. In the (post)semantics presented above—futurism—every sentence has exactly one of the two truth-values and future contingents can be true. I do not want to say that an actualist is forced to adopt this solution; they can decide to adopt one of the (post)semantics presented earlier and refuse to use the notion of the world of context instead. However, I believe this is a decision of a semantic, and not metaphysical, nature.

One should mind that when choosing one of the histories as the history of context, I indicated the history that "accurately represents the way the world was, is and will be". Consequently, in order to establish the truth-value of the sentence uttered in the given context, I implicitly referred to the past and future states of the world. Actualism guarantees that at every moment of the time, there is (was, will be) one such state. However, to use this state to our needs, we have to assume that we can refer to it when establishing the truth-value of an expression. I call this assumption "semantic transtemporalism". According to this statement, the truth-value of the sentence "At the time t, φ" assessed at the time t' depends on the way the world is (was, will be) at the time t, not the way it is at the time t'.

I believe the subject of the famous conflict between Kotarbiński (1913) and Leśniewski (1913) was in fact the question of justification of transtemporalism. Kotarbiński rejects this idea while Leśniewski defends it. Kotarbiński seems to have been swayed by Leśniewski's arguments but his way of thinking about the relation between truth and time was continued by Łukasiewicz.[11] Łukasiewicz persistently stood by localism, arguing that in order for the statement "At the time t' the coin lands tails up"

[11] It is not certain if Kotarbiński inspired Łukasiewicz in this matter or just on the contrary (see Woleński, 1990).

to be true at the time t, there must be conditions at the time t that *decide* that the coin lands tails up at the time t'. If t' is later than t, these conditions may be e. g. the angle or the force of the coin toss, as long as they combined guarantee the coin's landing tails up. If t' is earlier than t, the conditions are the traces left by the coin landing tails up (e. g. memories). If at the time t there are no conditions that guarantee the truth or falsity of the given sentence, it cannot assume any of the classic truth-values. This view is expressed by Łukasiewicz as early as 1922 (see Łukasiewicz, 1961, p. 122) and repeated by him in an almost unchanged form not long before his death (see Łukasiewicz, 1957, pp. 154–155).

This is, however, not the only way of thinking on the relations between truth and time. One can argue, in accordance with Leśniewski, that the truth-value of the sentence "At the moment t' the coin lands tails up" at the moment t should depend on what the state of the coin was or will be at the moment t' and not on the state the coin is in at the moment t. Not wanting to delve into the discussion of advantages and disadvantages of the two approaches here, I will only stress that I do not think transtemporalism should be in the losing position here.

Summing up, the traditional solution to the index initialisation problem is not excluded even for the tempo-modal semantics modelling indeterministic situations. However, a condition of applying this solution is to assume the metaphysical actualism and semantic transtemporalism. These are real commitments that one should be aware of. Nevertheless, I believe that when classical logic and a natural analysis of tempo-modal language is at stake, adopting these views is not too high a price.

REFERENCES

Adams, R. M. (1974). Theories of Actuality. *Noûs*, *8*(3), 211–231.

Belnap, N. (2002). Double Time References: Speech-Act Reports as Modalities in an Indeterminist Setting. In: D. Wolter, H. Wansing, M. de Rijke, M. Zakharyaschev (Eds.), *Advances in Modal Logic* (vol. 3, pp. 37–58). Singapore: World Scientific.

Belnap, N., Perloff, M., Xu, M. (2001). *Facing the Future: Agents and Choices in Our Indeterministic World*, New York: Oxford University Press.

van Fraassen, B. C. (1966). Singular Terms, Truth-Value Gaps, and Free Logic. *Journal of Philosophy*, *63*(17), 481–495.

Kaplan, D. (1989). Afterthoughts. In: J. Almong, J. Perry, H. Wettstein (Eds.), *Themes from Kaplan* (pp. 565–614). New York: Oxford University Press.

King, J. C. (2003). Tense, Modality, and Semantic Values. *Philosophical Perspectives*, *17*(1), 195–246.

Kotarbiński, T. (1913). Zagadnienie istnienia przyszłości [„The Problem of the Existence of the Future]. *Przegląd Filozoficzny*, *16*(1), 74–92.

Kripke, S. (1980). *Naming and Necessity*. Cambridge, Mass.: Harvard University Press.

Leśniewski, S. (1913). Czy prawda jest tylko wieczna czy też odwieczna [Is Truth Only Eternal or Both Eternal and Sempiternal?]. *Nowe Tory*, *10*, 493–528.

Lewis, D. (1970). Anselm and Actuality. *Noûs*, *4*(2), 175–188.

Lewis, D. (1986). *On the Plurality of Worlds*. Oxford: Blackwell Publishers.

Łukasiewicz, J. (1957). *Aristotle's Syllogistic from the Standpoint of Modern Formal Logic* (2nd edition). Oxford: Oxford University Press.

Łukasiewicz, J. (1961). O determinizmie [On Determinism]. In: J. Słupecki (Ed.), *Z zagadnień logiki i filozofii* (pp. 114–126). Warszawa: PWN.

MacFarlane, J. (2003). Future Contingents and Relative Truth. *The Philosophical Quarterly*, *53*(212), 321–336.

MacFarlane, J. (2008). Truth in the Garden of Forking Paths. In: M. García-Carpintero, M. Kölbel (Eds.), *Relative Truth* (81–102). Oxford: Oxford University Press.

MacFarlane, J. (2014). *Assessment Sensitivity: Relative Truth and Its Applications*. Oxford: Clarendon Press.

Plantinga, A. (1974). *The Nature of Necessity*. New York: Oxford University Press.

Ploug, T., Øhrstrøm, P. (2012). Branching Time, Indeterminism and Tense Logic. *Synthese*, *188*(3), 367–379.

Prior, A. (1966). Postulates for Tense-Logic. *American Philosophical Quarterly* 3(2), 153–161.

Prior, A. (1967). *Past, Present and Future*. Oxford: Clarendon Press.

Stalnaker, R. (1976). Possible Worlds. *Noûs*, *10*(1), 65–75.

Thomason, R. H. (1970). Indeterminist Time and Truth-Value Gaps. *Theoria*, *36*(3), 264–281.

Thomason, R. H. (1984). Combinations of Tense and Modality. In: D. Gabbay, F. Guenthner (Eds.), *Handbook of Philosophical Logic* (vol. 2, pp. 135–165). Dordrecht: D. Reidel Publishing Company.

Twardowski, K. (1900). O tzw. prawdach względnych [On the So-Called Relative Truths]. In: *Księga Pamiątkowa Uniwersytetu Lwowskiego ku uczczeniu pięćsetnej rocznicy Fundacji Jagiellońskiej* (pp. 64–93). Lviv: University of Lviv.

Wawer, J. (2014). The Truth about the Future. *Erkenntnis*, *79*(3), 365–401.

Wawer, J. (2016). *Branching Time and the Semantics of Future Contingents* (doctoral dissertation). Kraków: Jagiellonian University.

Woleński, J. (1990). Kotarbiński, Many-Valued Logic, and Truth. In: J. Woleński (Ed.), *Kotarbiński: Logic, Semantics and Ontology* (191–198), Dordrecht: Kluwer Academic Publishers.

Originally published as "Problem ustalania indeksu w semantyce temporalno-modalnej". *Studia Semiotyczne*, *31*(1), 109–130, DOI: 10.26333/sts.xxxi1.07. Translated by Agnieszka Przybyła-Wilkin.

Studia Semiotyczne—English Supplement, t. XXX (2019), s. 43–59
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/stsen.xxx.03
Article

Maciej Sendłak*

ABOUT THE BASIS FOR THE DEBATE OF COUNTERPOSSIBLES

Summary: According to the most popular (so-called "orthodox") theories, counterfactuals with impossible antecedents are vacuously true. Critiques of this view argue that contrary to this, we tend to consider only some of them true and others to be false. In his recent paper (*Counterpossibles*) Timothy Williamson has ingeniously explained the motivations for the orthodox view and argued that although there are some heuristic reasons that may suggest the plausibility of the unorthodox view, they are fallible. The most important of Williamson's arguments is that the unorthodox interpretation is inconsistent with the heuristic assumption that supposedly motivates this very view. The aim of this paper is to consider Williamson's critique and to support the unorthodox approach towards counterpossibles. In order to do so, we argue in favor of the modified version of the heuristic assumption.

Keywords: counterfactuals, counterpossibles, possible worlds semantics, methodology, Timothy Williamson.

The subject of this paper is the debate over truth-values of counterpossibles, i.e., subjunctive conditionals, the antecedents of which express

* University of Warsaaw, Institute of Philosophy. E-mail: maciej.sendlak@gmail.com. ORCID: 0000-0002-0539-5924.

impossibility.[1] On one side of this debate are advocates of so-called orthodoxy, which tracks back to the works of Robert Stalanker (1968) and David Lewis (1973), and which nowadays is defended by Timothy Williamson (2007; 2016b; 2018). Orthodoxy has it that every counterpossible is true. On the opposite side, there are advocates of unorthodoxy, who argue in favor of the thesis that some counterpossibles are false (Yagisawa, 1988; Nolan, 1997; Priest, 2009; Brogaard & Salerno, 2013).

The main aim of this paper is to reply to two of orthodoxy's arguments against motivations for the unorthodox approach (Williamson, 2016b; 2018). The first one has it that intuitions that underpin unorthodoxy are in tension with commonly accepted rules of counterfactuals. The second one aims to highlight a misunderstanding of the orthodox approach. Being an advocate of orthodoxy, Williamson argues that this misunderstanding results in an implausible characterization of this approach. Both arguments are meant to provide reasons for which unorthodoxy may be considered an implausible view.

I aim to look closely at those charges and to refute them. Firstly, I am going to argue in favor of the consistency of the unorthodoxy and the main rules of counterfactuals. Further, the question that Williamson considers to be based on a misunderstanding of orthodoxy will be revised. I believe that this will allow the justification of unorthodoxy.

Two aspects of the debate should be stressed right away. The subject of the debate is the truth-value of counterpossibles. Some critics of orthodoxy suggest that the vacuous truth of counterpossibles entails a lack of their semantic informativeness or that their meaning is independent of the consequent (Brogaard & Salerno, 2013). It is debatable whether orthodoxy entails this. This will not be a subject of this paper, for I am going to focus merely on the explicit thesis of orthodoxy, according to which every counterpossibles is vacuously true.

Secondly, my aim is not to argue in favor of either the inconsistency of orthodoxy or its implausibility due to the thesis of the vacuous truth of every counterpossible. It should be noted that some advocates of this approach try to provide an alternative explanation of the common intuition that some counterpossible are false. This is often done by moving the burden of the problem from semantics into pragmatics. Accordingly, it is claimed that while every counterpossible is vacuously true, there are good

[1] This material is based on the work supported by National Science Centre (NCN), Poland (Grant No.2016/20/S/HS1/00125).

pragmatic reasons for which we do assert some of them and do not assert others (Emery & Hill, 2017). While this is an exciting proposal and focusing on the pragmatic aspects of counterfactuals has a long tradition,[2] analysis of this goes beyond the aim of this paper. This is because more than the efficiency of the orthodox approach, I am interested in arguing in favor of the thesis that unorthodoxy is a consistent, well-motivated, alternative to orthodoxy, worth further development. After all, the lack of consistency is what unorthodoxy has been charged with by Timothy Williamson.

In order to do so, I shall begin with a rough characteristic of what counterfactuals are and what are the motivations that underpin both the orthodoxy and the unorthodoxy. After this, I shall focus on the argument that is meant to prove the inconsistency of unorthodoxy. Further, the question of a misunderstanding of orthodoxy will be reconsidered. The last part is devoted to the methodological aspect of the debate.

Counterfactuals

Counterfactuals are complex propositions that are often expressed as "If it had been the case that A, then it would be the case that C" ($A > C$), where A (antecedent), and C (consequent) are propositions, e.g.:

(1) "If the match had been scratched, it would have lighted."

(2) "If there had been no email controversy, Hillary Clinton would have won the election."

(3) "If Christopher Columbus had reached the place he was planning to reach in 1492, he would have arrived in Japan."

By the use of this kind of proposition, we indicate an essential connection between what is expressed by the antecedent and the consequent. We refer to them both in everyday life as well as in scientific discourses. They are considered to be an inherent aspect of gaining and transferring knowledge, expressing our beliefs, opinions, and attitudes, and stimulate our behavior (Edgington, 1995; Bennett, 2003; Williamson, 2016a).

It seems that one of the reasons for which we consider counterfactuals to have such importance for our intellectual life is that we ascribe them different truth-values. While we tend to consider (1) true, (2) is false.

[2] See the works of Grice (1975) and Jackson (1988).

Even though the claim that counterfactuals have different truth-values is close to banality, providing the proper truth criteria for such complex expressions is hardly a trivial endeavor. It should be noted that this is not the question of whether Columbus was planning to arrive in Japan, or of whether the email controversy was the only reason for which Clinton lost to Trump. While these may have some importance for the analysis of counterfactuals in general, the main issue is to provide a semantic criterion of truth-value for complex propositions such as (1–3).

Possible Worlds Semantics

The most popular analysis of counterfactuals is the one provided in terms of possible worlds semantics. This has it that sentences that contain modal operators of possibility—"it is possible that *p*" (or "it could be the case that *p*")—should be understood as ones that state that there is a possible world where *p* is the case. It is claimed that each sentence of the form "it is possible that *p*" is true if and only if there is a world (actual or merely possible) where *p* is the case. Thus, "Christopher Columbus could have reached Japan in 1492" should be interpreted as one which states that there is a possible world, where Christopher Columbus did reach Japan in 1492. Likewise, sentences that contain a modal operator of necessity, e.g., "It is necessary that *p*" (or "It has to be the case that *p*") are true if and only if in every possible world it is the case that *p*. Thus, "It is necessary that 2+2=4" is true because in every possible world, it is the case that 2+2=4. If it had been otherwise, i.e., if there had been a possible world where 2+2 does not equal 4, then we would have to admit the truth of "It is possible that 2+2 does not equal 4."

Possible worlds semantics, by providing an analysis of modality, became an attractive model for the analysis of counterfactuals. Based on this, the two very similar approaches of Robert Stalnaker (1968) and David Lewis (1973) have been proposed. According to these, $A > C$ is true in the actual world if and only if either:

(i) there is no possible world, where A is the case

or

(ii) there is a possible world w_1, where A and C are the case, and this world is more similar to the actual world than any possible world w_2, where A is the case, but C is not.

In virtue of the above, "If the match had been scratched, it would have lighted" is true because there is a world where the match is scratched and where it lights, and this world is more similar to the actual world than one where even though the match has been scratched, it does not light.

While possible worlds semantic is the most popular analysis of counterfactuals, it is not problem-free. One of these problems is somehow similar to that of paradoxes of material implication. As condition (i) has it, every counterfactual, which contain an impossible antecedent, is true. Thus, each of the below is true:

(4) "If there had been a round square, geometry would be different to what it actually is."

(5) "If there had been a round square, geometry would be the same as it actually is."

(6) "If it had been raining and not raining at the same time, some contradictions would be true."

(7) "(Even) if it had been raining and not raining at the same time, no contradictions would be true."

(8) "If whales were fish, they would have gills."

(9) "If whales were fish, they would not have gills."

Due to the impossibility of the antecedents (mathematical, logical, and metaphysical respectively) of (4–9), each of these is true.[3] After all, each of them satisfies the condition (i). Since the truth of (4–9) does not depend upon consequences, they are considered to be vacuously true. This means that these are true regardless of the consequents.

[3] This shows that impossibility is not restricted to merely logical impossibility, which is usually of the form of the conjunction of two opposite propositions, p and $\neg p$ (e.g., antecedents of (6) and (7)). It is claimed that an impossible state of affairs is a state that is realized in no possible worlds. Thus, if one admits that beyond logical truths the truths of mathematics and metaphysics are necessary, the antecedents of (4), (5), (7), and (9) also express impossibilities.

The above consequence seems to go against common intuitions. While we tend to consider (4), (6), and (8) true, they are not vacuously so. This partly depends upon the fact that we consider (5), (7), and (9) false. After all, the fact that no square is round is grounded in the laws of geometry, the truth of propositions of the form p and $\neg p$ is a contradiction, and one of the essential features of fish is that they have gills. Because of this, we are justified in expecting that an adequate analysis of counterfactuals will take these data into consideration and provide analysis, which would explain the falseness of expressions such as (5).

Philosophers who find this convincing argue in favor of a modification of possible worlds semantic analysis which is based on extending the domain of worlds to include impossible worlds, i.e., worlds where what is impossible in the actual world, is true. In virtue of this, some worlds contain round squares, true contradictions or whales that are fish. This results in a modified truth criterion of counterfactuals, which has it that $A{>}C$ is true if and only if there is a possible or impossible world w_1, where A and C are the case, and this world is more similar to the actual world than any possible world w_2, where A is the case, but C is not the case.

While this modification does justice to common intuitions about the falseness of some counterpossibles, it raises questions about the logical and metaphysical nature of worlds.[4] Even though this is a highly interesting issue, the plausibility of considering this is based on the assumption that the mentioned modification is justified in the first place. This assumption, however, is often questioned (Lewis, 1986, p. 7; Stalnaker, 1996). Among a number of arguments against belief in an impossible world, one aims to show that unorthodoxy on counterpossibles results in inconsistency (Williamson, 2018). Before going into details of this charge, I shall explicate the orthodox view.

Orthodoxy

The starting point of orthodoxy—as Williamson argues—is the fact that in virtue of intensional semantics every counterfactual with an impossible antecedent has the same intension, and hence the same truth-value.[5]

[4] See, e.g., (Berto, 2013).

[5] This is because the orthodoxy's domain of worlds does not include impossible worlds, which could represent various impossibilities.

This does not yet prejudge the question of whether each and every counterpossible is true or false. The additional assumption is that every counterfactual the consequent of which is a mere repetition of the antecedent (e.g., $A>A$) is true. This should not be controversial, for if there is a proposition of which we can be sure of its truth, $A>A$ seems to be the right candidate. Since this is true regardless of whether "A" expresses possibility or impossibility, the mentioned assumption applied to counterfactuals of possible antecedent ("If Christopher Columbus had reached the place he was planning to reach in 1492, he would have reached the place he was planning to reach in 1492") as well as to counterpossible ("If there had been a round square, there would have been a round square"). Thus, if one agrees that each counterpossible has the same intension, and that each "$A>A$" is true, every counterpossible is true (Williamson, 2018, p. 1).

An advocate of unorthodoxy could argue that one of the reasons for which we assume $A>A$ to be always true is that the negation of this, i.e., $A>\neg A$ is always false. After all, even if one has no knowledge with respect to A, one may assume that $\neg A$ is inconsistent with it and that it is impossible for both A and $\neg A$ to be true. Thus, the reason for a belief in the necessary truth of $A>A$ ("If Christopher Columbus had reached the place he was planning to reach in 1492, he would have reached the place he was planning to reach in 1492", "If there had been a round square, there would have been a round square") is indirectly a reason for a belief in the falseness of $A>\neg A$ ("If Christopher Columbus had reached the place he was planning to reach in 1492, he would not have reached the place he was planning to reach in 1492", "If there had been a round square, there would have been no round square"). This may suggest that the justification for the truth of $A>A$ is also a justification for the falseness of $A>\neg A$.

Contrary to the above, advocates of orthodoxy argue in favor of the thesis which has it that if A expresses impossibility, both "$A>A$" and "$A>\neg A$" are true. As Williamson argues, this is partly grounded in the commonly accepted principle that counterfactuals distribute over conjunction in the consequent: $((A>C)\ \wedge\ (A>B))\equiv(A>(C\ \wedge\ B))$. In virtue of this principle, the truth of $A>A$ and $A>\neg A$ result in the truth of $A>(A\ \wedge\ \neg A)$. While acceptance of this may raise some doubts, this merely shows that if the consequent of a given A is a contradiction, and if no contradiction is possible, the mentioned antecedent is not possible either (Williamson, 2018, p. 3). Thus, the acceptance of the truth of $A>A$ and $A>\neg A$ is grounded in the impossibility of A. In other cases, i.e., those

where A is possible, the truth $A>A$ entails the falseness of $A>\neg A$ (Stalnaker, 1968, p. 106).

The reasoning mentioned above is a justification for a belief in the vacuous truth of counterpossibles rather than criticism of unorthodoxy. Since this heavily relies on the assumption of the nonexistence of impossible worlds, the extension of the worlds' domain by introducing impossible worlds, would result in the situation where we could choose between two alternatives—orthodoxy and unorthodoxy. This would be a real choice only if both alternatives were consistent approaches. In this respect, Williamson charged unorthodoxy with being inconsistent. The mentioned inconsistency is meant to be grounded in the motivation for the belief in non-vacuous counterpossibles. This is the subject of the following section.

Misleading Heuristics

Considering the popularity of the orthodoxy, one may raise a question about the explanation of the common intuition which has it that some counterpossibles are false. Williamson sees the source of this intuition in what he calls heuristics, which is reflected in one of two expressions:

> (HCC) Given that C is inconsistent with D, treat $A>C$ as inconsistent with $A>D$.

or

> (HCC*) If you accept one of $A>C$ and a $A>\neg C$, reject the other. (Williamson, 2018, p. 8)

As Williamson argues, the belief in the plausibility of the above is what is meant to justify the unorthodoxy on counterpossibles. Thus, in virtue of either (HCC) or (HCC*), the truth of $A>A$ should result in the falseness of $A>\neg A$. This—advocates of unorthodoxy seem to claim—gives an accurate picture of the way in which we use counterfactuals with possible as well as those with impossible antecedents.

Contrary to this, it is argued that while in many cases, the use of the above-mentioned heuristics is justified, they do not apply unrestrictedly. A counterexample to this is a counterfactual with an impossible antecedent. As has been shown previously, an advocate of orthodoxy argues that in such cases, both $A>C$ and $A>\neg C$ are true. Thus (HCC) and (HCC*) apply to only those cases where the antecedent expresses possibility (Wil-

liamson, 2018, p. 9). In other cases, to rely on the heuristics results in a consequence that is inconsistent with the orthodoxy, i.e., the claim that some counterpossibles are false.

The above observation focuses on the relation between the orthodoxy and heuristics and shows why—in virtue of the former—the unrestricted acceptance of the latter is implausible. This, however, allows for an alternative interpretation. Namely, one according to which the thesis of orthodoxy contradicts the common phenomena expressed by (HCC) or (HCC*), so one should lean towards unorthodoxy. This would be justified if advocates of unorthodoxy could apply heuristics in an unrestricted way. As Williamson argues, this is not the case, which is meant to be shown by two counterpossibles:

a) $(A \wedge \neg A) > A$
b) $(A \wedge \neg A) > \neg A$

In virtue of (HCC) one should admit that the truth of (a) results in the falseness of (b). This, however, is problematic for at least three reasons. First of all, this would require rejecting one of the commonly accepted assumptions about counterfactuals, which has it that if an antecedent is a conjunction, then each conjunct of this is a consequent of this counterfactual, i.e. $(A \wedge B) > A$ and $(A \wedge B) > B$. Secondly, acceptance of only one of (a) and (b) contradicts the principle of counterfactual distribution over conjunction in the consequent. After all, since both (a) and (b) have the same antecedent, one should conclude (c): $(A \wedge \neg A) > (A \wedge \neg A)$. Finally, since (c) is an example of a counterfactual of the form $A > A$, the falseness of (c) goes against the initial assumption about the truth of every counterfactual of the form $A > A$. Thus, the consequences of the heuristics which meant to justify the unorthodoxy are incompatible with the general assumptions about counterfactuals (Williamson, 2018, p. 8).

In virtue of the above, an advocate of the unorthodoxy finds herself in a highly problematic situation. In order to defend this approach, one would have either to give up all of the three mentioned assumptions about counterfactuals or to modify the heuristics. I am going to argue in favor of the second option. Before doing so, however, it is worth mentioning what Williamson considers to be the misunderstanding of orthodoxy, i.e., the claim that an advocate of orthodoxy believes that the consequents of a counterpossible play no role when it comes to determining the

truth-value of a given counterpossible. This charge has been formulated by Beritt Brogaard and Joe Salerno:

> Counterpossibles are trivial on the standard account. By "trivial", we mean vacuously true and semantically uninformative. Counterpossibles are vacuously true in that they are always true; an impossibility counterfactually implies anything you like. And relatedly, they are uninformative in the sense that the consequent of a counterpossible makes no contribution to the truth-value, meaning or our understanding of the whole. (Brogaard & Salerno, 2013, p. 642)

The problem that Brogaard and Salerno pointed out is often considered an indirect motivation for rejecting the orthodoxy in favor of the unorthodoxy. According to Williamson, the charge is based on a misinterpretation of the first of these (Williamson, 2018, p. 4–5).

A Consequent of a Counterpossible

Williamson's argument is of the form of a reduction ad absurdum, and the crucial part of it is an analogy with other types of vacuously true counterfactuals, i.e., counterfactuals with necessarily true consequents. In virtue of this, it is claimed that if advocates of the orthodoxy claimed that the consequent of a counterpossible played no role in its truth-value, then the vacuous truth of a counterfactual with a necessarily true consequent would be independent of its antecedent. This would allow for a particular type of counterfactual, namely one which has an impossible antecedent and a necessarily true consequent:

(10) "If 6 were prime, 35 would be composite" (Williamson, 2018, p. 5).

Following the criticism of the orthodoxy—Williamson claims—one would have to admit that both the antecedent and the consequent of (10) have no contribution to the truth-value of this counterfactual. This, however, is implausible for without an antecedent and a consequent what is left is a bare form of the counterfactual sentence, which cannot give a truth-value on its own. If this is the consequence of the argument, then it is misleading for none of the advocates of orthodoxy would like to hold such a ridicules thesis (Williamson, 2018, p. 5).

If Williamson is right, the critique of orthodoxy should either argue that the mentioned "ridiculous thesis" indeed is a consequence of the or-

thodoxy or point out that this thesis is not a consequence of Brogaard and Salerno's charge. Choosing the latter option, I am going to argue that there is no need to believe that the mentioned charge results in ascribing to the orthodoxy the view that (10) is true in virtue of being a counterfactual.

What is key for Williamson's analysis is the question of what is the bare form of the counterfactual sentence. If we assume that the bare form of disjunction is the expression of the form $p \vee q$, then the bare form of the counterfactual sentence is $A > C$. The mere form does not allow for the determination of the truth-value of a counterfactual, which is reflected in the fact that philosophers of conditionals provide additional truth-conditions.[6] Likewise, the bare form of disjunction does not determine the truth value of $p \vee q$. While it is difficult to agree that the mere structure of (10) determines the truth value of it, one may question whether this is a consequence of the charge of Brogaard and Salerno. It seems that there are two reasons to believe that the claim that the consequent of counterpossibles does not contribute to the truth-value of the whole does not entail the thesis that (10) is true only in virtue of being a counterfactual.

The first reason is that if the claim mentioned above had been a consequent of Brogaard and Salerno's charge, the charge would have it that, in virtue of the orthodoxy, counterpossibles such as (8*) "If whales were fish, C" are vacuously true. This, however, would change the original subject of the charge, for this would be a problem of the vacuous truth of not well-formed formulas. This is due to the assumption that the counterfactual is a logical connective of two sentential arguments (A and C). Thus, in order to estimate the truth-value of it, both arguments should be satisfied by sentences. (8*) does not satisfy this for it contains one sentence and one sentential variable.[7]

While the belief in the truth of (8*) is controversial, this is not the aim of the original criticism of the orthodoxy. The aim is the thesis that regardless of what C is substituted by (8*) will be vacuously true. In this

[6] Williamson did not write explicitly what he means by "the bare form" of (10). Thus, one may raise doubts about whether the proposed "$A > C$" is actually the bare form of a counterfactual, for while this may represent the structure of (10), this does not reflect the modal status of its antecedent and the consequent.

[7] Based on the analogy to the bare form of disjunction, for every disjunction, where one of the disjuncts is "2+2=4" is true, this does not mean that "2+2=4 or p" (or "2+2=4 $\vee$ p") is true. After all, these are not well-formed formulas.

sense, a consequent of a counterpossible makes no contribution to the truth-value of the whole. Thus, one may question whether the consequence of Brogaard and Salerno's charge is the thesis that orthodoxy has it that what makes (10) true is the fact that is has the structure of $A > C$.

Secondly, if one asks an advocate of orthodoxy for motivations to believe in the vacuous truth of (4–9), she would say that this is so due to the impossibility of their antecedents. If we asked what makes the sentences "Even if Christopher Columbus had reached the place he was planning to reach in 1492, 36 would be composite" true, she would say that this is due to the necessary truth of the consequent. Both of these conditions—in virtue of orthodoxy—are sufficient to believe in the vacuous truth of the mentioned counterfactuals. Likewise, the truth of (10) is not grounded in the fact that the antecedent is impossible, and the fact that the consequent is necessarily true. What—in virtue of orthodoxy—makes (10) true is rather the fact that this satisfies a disjunction of conditions: a counterfactual is true whenever its antecedent is impossible, or the consequent is necessarily true. In the first case, the consequent plays no role in evaluation, in the second, the antecedent does not contribute to the truth-value.

This shows that contrary to what Williamson suggests, the criticism of orthodoxy does not have to entail the above-mentioned ridiculous thesis that (10) is true because of its structure. Nevertheless, the acceptance of orthodoxy results in the consequence that the impossibility of the antecedent determines the truth-value of the counterfactual. Thus, the consequent of a counterpossible (its meaning, modal status, or truth-value) makes no contribution to the truth-value of the whole expression.

Heuristics modified

The above allows us to move back to the question of heuristics. Timothy Williamson argues that the unrestricted acceptance of these is equally problematic for an advocate of orthodoxy as it is for the critiques of this approach. Thus, one should not consider them as a plausible motivation for rejecting orthodoxy in favor of orthodoxy. This is so due to the incompatibility of heuristics and the above-mentioned three principles that were meant to regulate the use of counterfactuals in general. In virtue of this, it is worth considering whether it is possible to provide such an alternative formulation of heuristics that on the one hand would justify the intuitions of different truth-values of counterfactuals (of possible or im-

possible antecedents), and on the other hand would not be in tension with the truth of (a) and (b).

It seems that the source of the incompatibility of (HCC*) and the truth of (a) and (b) is that while the consequents of (a) and (b) are incompatible with each other, each of them is compatible with the antecedent $A \wedge \neg A$. After all, if the antecedent is of the form of conjunction, the consequent can be any of the conjuncts. Likewise, in the case of orthodoxy, where the inconsistency of A and $\neg A$ does not preclude making them both consistent consequences of the impossible antecedent. This shows how crucial for the evaluation of counterfactuals is the antecedent and might be a good starting point for a reformulation of heuristics. If one would like to express orthodoxy in terms of heuristics one could say that "Assuming the possibility of A, if you accept one of $A > C$ and $A > \neg C$, reject the other." This seems to reflect the way in which advocates of orthodoxy think about counterfactuals. At the same time, this shows that the tension between $A > C$ and $A > \neg C$ arises only if A is possible. Thus, one can formulate orthodoxy's heuristics, which has it that:

> (HCC**) "If A does not allow for the simultaneous acceptance of them both, if you accept one of $A > C$ and $A > \neg C$, reject the other."

Somehow similar heuristics apply to the unorthodoxy as well. The difference here lies in the fact that the impossibility of A is not a sufficient condition for the acceptance of both $A > C$ and $A > \neg C$. This, however, does not have to be a deal-breaker, for (HCC**) says nothing about what exact conditions A has to satisfy. Thus, (HCC**) can be easily accepted by unorthodoxy to expresses the motivation for this view. This can be done by claiming that while $(A \wedge \neg A) > A$ and $(A \wedge \neg A) > \neg A$ have opposite consequences, both are true due to the fact that both consequences are compatible with the antecedent. Thus, in this particular case, the antecedent does allow for the simultaneous acceptance of both counterfactuals.

It should be noted that regardless of whether one favors orthodoxy or unorthodoxy, the majority of counterfactuals satisfy (HCC*). Nevertheless, there are also examples of pairs of counterfactuals with opposite consequents, which makes it implausible to use the mentioned heuristics in an unrestricted way. This makes (HCC*) misleading and merely partly reflecting the way in which we use counterfactuals. The more accurate formulation of heuristics is (HCC**), which—contrary to (HCC) and

(HCC*)—does not have to be restricted to a particular type of counterfactuals. Moreover, this can be applied by both the orthodoxy and the unorthodoxy. Considering the lack of restriction in the application of (HCC**), there is a good reason to consider this not as a misleading heuristics, but rather as a normative rule, which expresses the relation between counterfactuals that have the same antecedents, but opposite consequents.

Nevertheless, if one accepts (HCC**), there is a question of why this is supposed to support the unorthodoxy analysis of counterpossibles. After all, this rule equally supports the orthodoxy, which suggests that this does not move us closer to the finding of an adequate approach towards counterpossibles. While this may be the case, one of the theoretical benefits of acceptance of (HCC**) is that this justifies the thesis of this paper, i.e., the consistency of the unorthodoxy motivation and other commonly accepted rules of counterfactuals.

The consistency of the unorthodoxy does not have to end the debate over an adequate analysis of counterfactuals. For—as Timothy Williamson claims—advocates of the unorthodoxy have to believe in impossible worlds, which (along with other assumptions of the unorthodoxy) results in implausible hybrid semantics. Compared to this, the unified orthodox approach seems to be more attractive (Williamson, 2016b). This leads to a consideration of methodological aspects of the debate over counterpossibles.

Methodological Aspects

Since the acceptance of (HCC**) is consistent both with the orthodoxy and the unorthodoxy, one may believe that the debate can be framed as a clash of intuitions with respect to the adequate analysis of counterfactuals. Thus, one faces two alternatives. The first one is a simple model, which—for the last decades—has been considered to be the default one, and which has it that every counterpossible is vacuously true. The alternative to this is a relatively new approach, which extend the worlds' domain by introducing impossible worlds, and which has it that some counterpossibles are false.

Considering their theoretical virtues, the two approaches highlight different methodological aspects. An advocate of orthodoxy points to the simplicity of her view, which is reflected in the simpler domain of the worlds. While simplicity is an essential theoretical virtue, this surely is

neither the only one, nor the most important.[8] This is due to the fact that the alternative's being less simple might be well motivated by its higher explanatory power. This condition is implicitly included in the principle of parsimony (so-called "Occam's Razor"), which has it that "entities should not be multiplied beyond the necessity." While the principle is one of the most popular, the vast majority of philosophers usually focus only on its first part and overlook the second part. For it is not the case that entities should not be multiplied at all, but instead they should not be multiplied beyond necessity. It is justifiable to consider the mentioned necessity to be an explanation of data that are the subject of a given theory. Thus, the principle of parsimony should be interpreted as one which has it that if two theories have the same explanatory power, one should favor the simpler one, i.e., the theory which postulates fewer entities, hypotheses, axioms, etc.

In virtue of this, while the orthodoxy is with some respects simpler, the complexity of the unorthodoxy's alternative has a good reason. This is the higher explanatory power, which is reflected in taking into consideration pre-theoretical intuitions of different truth-values of counterfactuals such as (4–9). Thus, the complexity of unorthodoxy does not have to be considered as a violation of the principle of parsimony. On the contrary, the entities that are in this case multiplied, are necessary for the explanation of the data.

This line of defense of unorthodoxy may be faced with the problem of officiousness. This problem arises when a theory is too sensitive when it comes to identifying data (Hitchcock & Sober, 2004). As philosophers who characterized this problem argue, we are often wrong when it comes to the recognition of what is the real data and what is merely "noise" in the data (Hitchcock & Sober, 2004, p. 10). In such cases, we are faced with the problem of wrong identification of what is meant to be explained by a given theory. Accordingly, our expectation of a theory to explain a given phenomenon is unjustified.

The inaccurate identification of data may lead to further complications. After all, if we consider what is merely noise to be real data, there is a risk of introducing unjustified changes in the original theory or simply rejecting the original theory in favor of the new one. This often happens because of a wrongly construed counterexample to the original theory

[8] It seems that some consider the parsimony to be merely a question of the aesthetic aspect of a given theory (Barcan Marcus, 1995, p. 199).

(Williamson, 2018). As Williamson claims, counterfactuals such as (4–9) can be considered such wrongly construed counterexamples. What makes them inadequate is that the intuition they are supposed to reflect is based on the (HCC), which is meant to be implausible.

If the main reason for which the unorthodoxy is implausible is meant to be due to (HCC) or (HCC*), an advocate of this approach might point to (HCC**). As I have argued, this seems to go along with the way in which we ascribe truth-values of counterfactuals. At the same time, this is general enough to be consistent with both orthodoxy and unorthodoxy. Importantly, this allows for an indication of the consistency of the latter.

REFERENCES

Barcan Marcus, R. (1995). *Modalities*. Oxford: Oxford University Press.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Berto, F. (2013). Impossible Worlds. In: E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from: http://plato.stanford.edu/entries/impossibleworlds

Brogaard, B., Salerno, J. (2013). Remarks on Counterpossibles. *Synthése*, *190*(4), 639–660.

Edgington, D. (1995). On Conditionals. *Mind*, *104*(414), 235–329.

Emery, N., Hill, C. (2017). Impossible Worlds and Metaphysical Explanation: Comments on Kment's "Modality and Explanatory Reasoning". *Analysis*, *77*(1), 134–148.

Grice, H. P. (1975). Logic and Conversation. In: P. Cole, J. Morgan (Eds.), *Syntax and Semantics, 3: Speech Acts* (41–58). New York: Academic Press

Hitchcock, C., Sober, E. (2004). Prediction Versus Accommodation and the Risk of Overfitting. *British Journal for the Philosophy of Science*, *55*(1), 1–34.

Jackson, F. (1988). *Conditionals*. Oxford: Blackwell.

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (1986). *On the Plurality of Worlds*. Oxford: Blackwell.

Nolan, D. (1997). Impossible Worlds: Modest Approach. *Notre Dame Journal of Formal Logic*, *38*(4), 535–572.

Priest, G. (2009). Conditionals: A Debate with Jackson. In: I. Ravenscroft (Ed.), *Minds, Worlds, and Conditionals: Themes from the Philosophy of Frank Jackson* (311–335). Oxford: Oxford University Press.

Stalnaker, R. (1968). A Theory of Conditionals. In: N. Rescher (Ed.), *Studies in Logical Theory* (98–112). Oxford: Blackwell.

Stalnaker, R. (1996). Impossibilities. *Philosophical Topics*, *24*(1), 193–204.

Williamson, T. (2007). *Philosophy of Philosophy*. Oxford: Oxford University Press.

Williamson, T. (2016a), Knowing by Imagining. In: P. Kung, A. Kind (Eds.), *Knowledge Through Imagination* (113–132). Oxford: Oxford University Press.

Williamson, T. (2016b). Counterpossibles in Metaphysics. In: F. Kroon (Ed.), *Philosophical Fictionalism*. Oxford: Oxford University Press.

Williamson, T. (2018). Counterpossibles. *Topoi*, *37*(3), 357–368.

Yagisawa, T. (1988). Beyond Possible Worlds. *Philosophical Studies*, *53*(2), 175–204.

Originally published as "U podstaw sporu o kontrmożliwe okresy warunkowe". *Studia Semiotyczne*, *31*(1), 131–151, DOI: 10.26333/sts.xxxi1.08. Translated by Maciej Sendłak.

Studia Semiotyczne—English Supplement, t. XXX (2019), s. 61–84
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/stsen.xxx.04
Article

Gabriela Besler*

GOTTLOB FREGE ON TRUTH DURING THE PERIOD OF THE TWO VOLUME EDITION OF *GRUNDGESETZE DER ARITHMETHIK* (1893–1903)

Summary: In 1893 and 1903, two volumes of the most important of Frege's works *Grundegezte der Arithmetik* were published. This period can be called the peak of Frege's logicism. Although the subject of truth in Frege's logical and philosophical works has been repeatedly investigated, there is a lack of studies on his view in this period, especially in Polish literature.

In this article, therefore, I carry out the following research task: to collect and order Frege's statements about truth during the period of publishing the two volumes of *Grundegezte der Arithmetik*.

I refer to the texts published during Frege's life, published posthumously and his correspondence. Particularly noteworthy are: *Grundegezte der Arithmetik* and the unpublished *Logik* (1897). That is why there are two separate sections dedicated to these two texts, whereas the discussions of truth from other texts are grouped thematically: the problem of antinomy, geometry, expression of generality, and others. The subject of truth appears there in relation to logic, philosophy of language, philosophy of logic, philosophy of mathematics and ontology.

Keywords: Gottlob Frege, truth, logic, truth-values, thought, logicism.

* University of Silesia in Katowice, Institute of Philosophy. E-mail: gabriela.besler@us.edu.pl. ORCID: 0000-0002-1843-5198.

Introduction

After number, truth is the second great subject in the philosophy of Gottlob Frege. In his over forty years of scientific activity, he repeatedly modified his understanding of it.[1] The changes, however, were not essential, but rather meant seeking the coherence of the proposed understanding of truth and specifying the original intuitions. I would like to add that Frege was always against the definition of truth as the correspondence of ideas [*Vorstellungs*] or a sentence [*Satz*] with reality.

In 1893 and 1903, two volumes of Frege's most important work *Grundegezte der Arithmetik* were published. It was a period[2] that could be called Frege's "peak of logicism", a period in which he believed in the success of his project and worked on its further development.

Although the subject of truth in Frege's logical and philosophical works has been repeatedly investigated, diachronically (Sluga, 2002) and synchronously (Burge, 2005; Dummett, 1981; Greimann, 2003a; 2003b; 2007), there is a lack of studies on his view in the peak period of the development of his logicism, especially in Polish literature, where only scattered comments can be found, focusing on the True as a truth-value, however, the topic is worth a major study.

In this article, therefore, I carry out the following research task: to collect and order Frege's statements about truth during the period of publishing the two volumes of *Grundegezte der Arithmetik*. I refer to the papers published during Frege's life, published posthumously and his correspondence. The subject of truth appears there in the context of logic, philosophy of language, philosophy of logic, philosophy of mathematics and ontology.

Frege's scientific achievements from 1893–1903 are very specific. Between the two volumes of *Grundegezte der Arithmetik* there appeared a small book on numbers in Schubert's approach, rarely mentioned as Frege's book (1899/1990). Of the remaining 29 papers 6 were published during Frege's life, 5 were neither published in his lifetime nor prepared

[1] For more on this subject see (Sluga, 2002; Besler, 2010, p. 189–201).

[2] Actually, the period lasted until June 16th in 1902, when Frege received the first letter from Russell (Russell, 1902/1976) informing him about the possibility of constructing an antinomy based on Frege's first book (Frege, 1879/1997). Next, Frege himself formulated an antinomy based on the logical system from *Grundgesetze der Arithmetik* (Frege, 1893; 1903). For more on this subject see (Besler, 2016).

by him for publication. The remaining 18 documents are letters, mostly of a scientific nature, written with great care and addressed to eminent scientists of that time: Giuseppe Peano (1858–1932), David Hilbert (1862–1943), Heinrich Liebmann (1847–1939), and Bertrand Russell (1872–1970).

Thus, a total of 32 documents were created by Frege in the period studied here, not all of which, however, refer to the subject of truth. Particularly noteworthy are: *Grundegezte der Arithmetik* (Frege, 1893/2009) and the unpublished *Logik* (Frege, 1897/1983). That is why there are two separate sections dedicated to these two texts, whereas the discussions of truth from other papers are grouped thematically: the problem of antinomy, geometry, expression of generality, and others.

There are a lot of words connected with truth in Frege's writing. Here is a list of the terminology typical for the period studied here, from the first volume of *Grundgesetze der Arithmetik*, along with the number of times each term is mentioned: the True [*Das Wahre*] (150), truth [*Wahrheit*] (112), truth-value [*Wahrheitswert*] (97), the False [*das Falsche*] (78), laws of being true [*Gesetze des Wahrseins*] (12). In addition, Frege often used an adjective (predicate) true/false [*wahr*/*falsch*].[3] The above terminology also occurs in other texts from the period examined here, with exceptions being indicated.

In the last position from the period examined here (Frege, 1903/2009a), the above words occur less frequently, and the wordings "the False" and "laws of being true" are not present at all. It does not mean that Frege had changed his point of view, but can be attributed to the fact that it is a different type of book. An appendix (Frege, 1903/2009b) was added to the book (Frege, 1903/2009a), and in it an attempt to improve the system after a difficulty formulated by Russell (1902/1976). From the above words appear there (as technical words) only (in one sentence alone): "the True", and "the False".

There are some words fundamentally connected with truth: value-range of a function [*Werthverlauf*], thought [*Gedanke*], contradiction [*Widerspruch*], declarative sentence [*Behauptungsatz*], judgement [*Urtheil*], science [*Wissenschaft*]. It is worth adding that in the paper *Logik* (Frege, 1897/1983) there is no expression "truth-value". Moreover, nowhere in the

[3] Frege's German terminology is translated into English in various ways. Here I rely on the solutions adopted by the editors of the new English translation of *Grundgesetze der Arithmetik* (Frege, 2016). I do not interfere in the translation of quotations. Often—for clarity—I give the original German words.

papers discussed here (or to be more precise, in any of Frege's documents) is there an expression "truth conditions" (*Wahrheitsbedingungen*) often invoked by analytical philosophers referring to Frege's idea of truth (see Dummett, 1981, p. 71; Besler, 2010, p. 76).

In the legacy of Frege, three papers have been found, which are treated as unfinished textbooks on logic: (Frege, 1879–1891/1983), (Frege, 1897/1983), (Frege, 1906/1983). The fourth textbook includes articles published as *Logische Untersuchungen*: (Frege, 1918–1919/1990a; 1918–1919/1990b; 1923/1990). In the last article, Frege presented his point of views on truth, thought, sense and reference, nature of logic, negation, and generality. This subject area corresponds to the subjects of his previous unfinished textbooks on logic (Frege, 1897/1983). In none of the above-mentioned documents is there Frege's logical notation, and their subject matter falls within the scope of philosophy of logic.

It is assumed that *Logik* was written in 1897, between the publication of the two volumes of *Grundgesetze der Arithmetik*. The central theme of *Logik* is truth, as substantially connected with logic.

It seems that dating this paper should not present any difficulties, because Frege gave the date in the sentence: "[...] at noon on 1st January 1897 by central Europe time" (Frege, 1897/1983, p. 135).[4] Moreover, German editors established that Frege mentioned:

1. Wilhelm Wundt's journal *Grundüge der physiologischen Psychologie*, which had appeared since 1874 (p. 144).
2. A review published in 1897 (p. 146).

However, one might be surprised by the similarity of many theses concerning truth and thought to the ones from a much later paper (Frege, 1918–1919/1990a). Here are some possible explanations for this situation, each involving a counterargument:

1. The text was written much later, and the date was not related to the date of writing it. Maybe it was meaningful to Frege for reasons unknown to us. Against this solution is his reference to the review from 1897.

[4] In the whole article the pages numer refer to English translations of Frege's papers. See References for details.

2. Frege did not change his views for twenty years or he returned to previously developed solutions. If so, then the concept of objective thought as something for the question of the True and the False appeared much earlier than the work from the retirement period. Against this solution is the lack of repetition of these theses in other writings from that period, including his letters.
3. Furthermore, there is a lack of the expression "truth-value" in this paper, which was crucial for the examined period.

For the purposes of this article, I assume, however, that *Logik* (Frege, 1897/1983) was written in 1897, between the two volumes of *Grundgesetze der Arithmetik*.

Grundgesetze der Arithmethik (1893)

The task of the first volume of *Grundgesetze der Arithmetik* and the subsequent planned volumes, of which only the second one appeared (Frege, 1903/2009b), was the presentation of arithmetic as developed logic (Frege, 1893/2009, p. VII[5]). Frege wrote there that logic deals with the laws of being true, unlike psychology, which is interested in laws of thought (p. XVI). In this context, the subject of truth appeared from the point of philosophy of language, and—along with the True, the False—as categories used in logic.

The philosophical aspect of truth is presented in *Vorwort*, one of two introductions to the first volume.[6] In the examined period, Frege's philosophy of language was already fully developed and well-established and he referred to his previous article (Frege, 1892/1990b).

He used philosophy of language to characterize truth. The basis was the distinction of (only) three types of linguistic expressions: a proposition [*Satz*], a proper name, and a predicate. Each of these expressions has its sense and the reference (understood as the "object" to which the expression referred).[7] The sense of a proposition is a thought, and its reference

[5] The page numbers referred to the canonical paging of this book, assumed also in (Frege, 2016).

[6] It is an issue for a separate investigation as to why Frege wrote two different introductions, one called *Vorwort*, the other *Einleitung*.

[7] For more on sematic categories in *Grundegezte der Arithmetik* see (Heck, 2010).

is one of two truth values:[8] the True or the False. All true (false) propositions refer to the same object, the True (the False). Here are some examples showing this point of view:

> The names "$2^2 = 4$" and "$3 > 2$" refer to the same truth-value, which I call for short the True. [...] The function $\xi^2 = 4$ can therefore only have two values, namely the True for the argument 2 and -2 and the False for every other argument. (p. 7)

On the basis of the above quotations it can be generally said that every true proposition is a proper name of an object, which is one of two truth-values, being the True in example above. And similarly with false propositions.

Frege arrived at the use of sense and reference in the context of truth from a different side. Content, as an element distinguished from the acknowledgment of the truth, was described by him as judgeable, and he distinguished two more elements (Frege, 1893/2009, p. X) in it:

1. Thought, which is the sense of proposition.
2. Truth-value, which is the reference of proposition.

From an historical point of view the expression truth-value proved to be the most important for his philosophy and logic, in fact for all logic in 20th century. He wrote: "I distinguish two truth-values: the True and the False." (Frege, 1893/2009, p. X)

The truth-value and the number [*Zahl*], but not cardinal number [*Anzahl*], were understood as objective, real, ideal objects. The objects were characterized by the fact that in their own name, meaning a proper name, "[...] they do not [...] carry argument place" (Frege, 1893/2009, p. 7).

It is necessary to add that functions (including propositional functions) do not have a truth-value, because as expressions with a variable they are incomplete. Functions "obtain" their truth-value only when they are completed by arguments. However, then, they are not functions any more. For Frege, a concept is "a function whose value is always a truth-value,

[8] The language of values was introduced into philosophy by Hermann Lotze (1817–1881) and Wilhelm Windelband (1848–1915). Frege was in contact with these academics. Windelband used the expression *Wahrheitswert*, Lotze—*Gedanke*, both of which differed from Frege's understanding of value in logic (Sluga, 2002, pp. 84–85; Besler, 2010, pp. 27–28, 73–81).

the True or the False" (p. 8), for example: the concept of red is actually a function "( ) is red", true for some arguments, false for others.

In a paper devoted to the comparison of his and Peano's notations Frege repeated the above-mentioned notions: "[...] all true sentences [*Sätzen*] mean the same thing, namely the True, and likewise all false sentences mean the same thing, namely the False" (Frege, 1896/1990, p. 240). "I use the word *Satz* in the sense of a combination of symbols whose sense is a thought and whose reference is a truth-value—either the True or the False" (Frege, 1896/1990, p. 242).

Incorrectly constructed propositions are treated as false in Frege's logic and his philosophy of language. (Frege, 1893/2009, p. 10; Frege 1896/1990, p. 230). He gave the following example. He introduced a sign for Sun Θ and using mathematical language wrote that the sign is greater than 2: "$\Theta > 2$". Frege called such a proposition false because Sun is not a number, however, the following proposition is true: "$(\Theta > 2) \supset (\Theta^2 > 2)$" (Frege, 1896/1990, p. 230). From the definition of the material implication we know that such a formula is true when predecessor and successor are false.

It is necessary to add that not every syntactically correct sentence possesses a truth-value. Frege pointed out two situations:

1. Subordinate clause in indirect speech. Generally, a thought is the sense of a proposition, however, in indirect speech the thought is treated as the reference of the subordinate clause (Frege, 1893/2009, p. X). Thus, the subordinate clause, as a part of indirect speech does not possess a truth-value.
2. Sentence with a proper name without reference like a sentence in poetry (Frege, 1896/1990, p. 227); such a rule was explicitly expressed in a later paper, however, in this period it is also valid (Frege 1897–1898/1983, p. 156).

Frege's views discussed above show that the True is essentially connected with his concept of thought. Actually, not only the True, but the False as well. Frege also wrote about false thoughts, giving the following examples: $0^2 = 4$; $1^2 = 4$; $3^2 = 4$ (1893/2009, p. 6).

According to Frege, the notion of thought, which supplements his categories of the Truth and False, is the meaning of the name of a certain logical value (Frege, 1893/2009, p. 7). Later, he even wrote about the "realm of thought" (Frege, 1918–1919/1990a), considering it an objective

reality, unchanging, guaranteeing the possibility of doing science, significantly connected with logic. In the context of logic he wrote: "[...] I express thoughts with my signs, it will be helpful to look at some of the easier cases in the table of more important theorems, to which a translation is appended" (Frege, 1893/2009, p. XI).

Truth is a notion that is also used in Frege's formal logic. For example, logical laws are called laws of being true (Frege, 1893/2009, p. XVI). Some of the logical laws served as the basic laws, not proved in Frege's system, one of them being the problematic law V.[9]

In the period of writing the two volumes of *Grundgesetze der Arithmetik*, the truth-values were used by Frege to determine the conditions for propositions constructed both of connectives and quantifiers to be true. Earlier in this context, Frege used the words: affirm [*bejahen*], deny [*verneinen*] (Frege, 1879/1997, p. 5),[10] instead of the True and the False.[11] There are the logical symbols that Frege characterized with a reference to the truth-value (I give them in Frege's order): judgement-stroke, horizontal-stroke, negation-stroke, equality-sign, quantifier-sign, conditional-sign.

He mentioned his first book (Frege, 1879/1997) and distinguished "two components in that whose external form is a declarative sentence:

1. Acknowledgement of truth.
2. The content, which is acknowledged as true" (Frege, 1893/2009, p. X).

The "acknowledgment of truth" is "marked" on a logical formula by attaching the so-called judgement-stroke and Frege described it as follows:

> We are therefore in need of another special sign in order to be able to assert something as true. To this end, I let the sign "$\vdash$" precede the name of the truth-value, in such a way, e.g., in $\vdash 2^2 = 4$ it is asserted that the square of 2 is 4. (Frege, 1893/2016, p. 9; Greimann, 2000)

[9] For more on this subject see the section *The Problem of Antinomy*.

[10] In the English edition p. 121. Apart from this example, page numbers are given from the English editions.

[11] It is worth adding, that Ernst Schröder (1841–1902), Charles S. Peirce and Frege are treated as originators of truth tables. However, Schröder used the expressions *es gilt*, *es gilt nicht* in this context (Marek, 1993, p. 10–11).

The necessity of this judgment-stroke is so obvious, natural and necessary to Frege that in a letter to Peano, with whom he corresponded during this period, he wrote:

> I have [..] the sign |, the judgement stroke, which serves to assert something as true. You have no corresponding sign, but you acknowledge the difference between the case where a thought is merely expressed without being put forward as true and the case where it is asserted. (Frege, 1896/1976, p. 185–186)

The so-called horizontal-stroke, is a sign of a one-place function from objects whose value is one of two truth-values (Frege, 1893/2009, p. 16–17). The value of this function is the True when its argument is the True. There are two other cases:

1. The False is the function's argument.
2. None of the truth-values is the function's argument, but for example, the number 2 (Frege, 1893/2009, p. 10).

Then the value of the function is the False.

The negation (written as a short stroke attached to the horizontal-stroke) is defined as the value of a false function for every argument and this function without a negation sign is true for every argument (p. 10).

An expression with the equality-sign refers to the True when expressions with the same logical value appear on both sides of the connective and to the False in any other case (p. 11).

The universal quantifier was written by Frege as a concavity in the content-stroke with the Gothic alphabet letter. He assumed that the formula "[...] refer[s] to the True if the value of the function $\Phi(\xi)$ is the True for every argument, and otherwise the False" (p. 12).

The conditional-sign (a sign for material implication) was written as a vertical stroke connecting two horizontal strokes and characterized as the False when the predecessor is the True and the successor is not the True (p. 26).

Frege also gives examples of functions whose value for every argument is the False:

1. The formula $\dot{\varepsilon}(\varepsilon = \neg\breve{\mathfrak{a}}\,\mathfrak{a} = \mathfrak{a})$ was read as the value-range of a function "it is denied that for every a, $a = a$" (p. 17).

2. Connecting one of the truth-values to a value-range with an equality-sign (see p. 17).

In Frege's logic, truth was also presented by logical entailment. In a paper written between 1899 and 1906 he wrote: "Truths [*Wahrheiten*] can be inferred in accordance with logical laws of inference. If a truth [*Wahrheit*] is given, it can be asked from what other truths its truth follows in accordance with logical laws of inference" (Frege, 1899–1906/1983, p. 168).

To sum up the topic of truth in *Grundgesetze der Arithmetik*, it can be said that it was both a philosophical notion and a useful "tool" for studying the truth-value of logical formulas, inferences and the characteristics of connectives or the quantifier. The True was expressed verbally or using the assertion-stroke.

Logik (1897): An Unfinished Textbook

This paper, unfinished and not published by Frege, is worthy of special attention for a number of reasons, including the method Frege used: referring to the ways of using the word "true" in ordinary language. He also pointed out words associated with the predicate "true", and words that do not have a significant relationship with it, although these expressions were used in an ordinary language. Next, he collected contexts in which the word "true" occurred and rejected misleading, improper usage. He compared the predicate "true" to other predicates (p. 126, 128)[12] and listed differences. The predicate "true" had nothing in common with the ideas [*Vorstellungen*], and was not "applicable to what is material" (p. 126).

Frege suggested setting limits on the valid applicability of the word "true". Although he did not explicitly state this position, it can be said on the basis of this and other papers that the predicate "true" refers to thoughts first and, sentences second, and in particular assertoric sentences [*Behautungssätzen*] (p. 126, 129). For sentences [*Sätzen*] are "a proper means of expression for a thought" (p. 126), and "a sense of the sentence is called a thought" (p. 126).

[12] All quotations from this section, unless otherwise stated, come from the paper mentioned in the heading.

In a natural language, "true" is also combined with ideas and experience, which Frege rejected as groundless. He also wrote that we do not need the word "true" to say that the idea of the Cologne cathedral agrees with reality. As the legitimate use of the word "true" he gave predicating it on a proposition like $2 + 3 = 5$ (p. 129). If, however, one speaks of an idea called true "[...] it is really a thought to which the predicate is ascribed" (p. 126).

Although truth is the goal for all science, logic is in a special way related to the predicate "true", like ethics to "good", aesthetics to "beautiful", physics to "heavy" and "warm", chemistry to "acid" and "alkaline" (p. 128). The word "true" specifies the goal of logic (p. 126).

The "true" and "beautiful" predicates, however, differ significantly. There may be a contradiction between propositions of logic, but "[a]esthetic judgements don't contradict one another" (p. 126). What is true—as Frege wrote—is "true in itself" (p. 126) and what is beautiful is not "beautiful in itself" (p. 126). In addition, the "true" predicate is not gradable, unlike "beautiful"—which can be graduated (p. 126).

For Frege, logic, like ethics, is the normative science based on the most general laws of truth (p. 128). Next, he wrote:

> Logic is concerned with the laws of truth, not with the laws of holding something to be true, not with the question of how men think, but with the question of how they must think if they are not to miss the truth. (p. 149)

That is why the laws of truth are contrasted with the laws of thinking and the laws of judging that psychology deals with (p. 145–146). Moreover, "[t]he laws of truth, like all thoughts, are always true if they are true at all" (p. 148).

In the unfinished textbook on logic, Frege clearly wrote about the indefinability of truth for the first time: "Truth is obviously something so primitive and simple that it is not possible to reduce it to anything still simpler" (p. 129). Therefore, he considered truth to be indefinable. In such cases one only has to "[...] to lead the reader or hearer, by means of hints, to understand the word as it is intended" (Frege, 1892/1990a, p. 183). In a later paper, from 1914, this activity would be called elucidation

[*Erläuterung*], distinguished from defining (Frege, 1914/1983, p. 207).[13] However, it seems strange because he described elucidation as a prescientific activity, beyond science, it was only its propedeutics.

Frege devoted a lot of space to the concept of thought, to which the predicate "true" fundamentally referred. Next, the predicate "true" referred to a declarative sentence.

In his concept of thought, first, truth (or falsity) is not a matter of recognition by one person or another. The objectivity of truth (or falsity) results from the "fixing" in objective thought. Frege believed that this guarantees objectivity in science. He wrote:

> [...] thoughts have [not] to be thought by us in order to be true. [...] Thoughts are independent of our thinking. A thought does not belong specially to the person who thinks it, as an idea does to the person who has it. [...] A contradiction between the assertion [*Behauptungen*] of different people would be impossible. (Frege 1897/1983, p. 127)

Next, thought is not mental. But if it were, then:

1. "[...] its truth could only consist in a relation to something external, and that this relation obtained would be a thought into the truth of which we could inquire" (p. 127).
2. Mathematical propositions would look as follows: "It has been observed that with many people certain ideas form themselves in association with the sentence '2 + 3 = 5'" (p. 134).

To sum up, I would like to emphasise the similarity between the unpublished *Logik* (1897) and *Der Gedanke* (1918–1919) published twenty years later; however, this topic needs further study.

Truth in Particular Contexts

Problem of Antinomy

Questions of truth, falsity and words connected with them occurred in the Frege—Russell correspondence. This exchange of letters referred

[13] In this paper, *Erläuterung* is translated as illustrative examples, however, in the literature in this context *elucidation* is used. See (Weiner, 2002; Besler, 2010, p. 148–149).

mainly to the problem of antinomy and the edition of the second volume of *Grundgesetze der Arithmetik* (Frege, 1903/2009a, with Frege, 1903/2009b).[14] Truth appears there in philosophical and logical contexts.

Frege tried to convince Russell on his philosophy of language, in which task he failed. Nevertheless, we have many clear passages related to philosophical solutions adopted for truth and falsity. They do not bring anything new to Frege's previous point of view, but they are worth special attention due to their precision and the unambiguity of the wording. Here are two examples, one from 1902, the other one written a year later:

> As you know I distinguish between the sense and the meaning [*Bedeutung*] of a sign, and I call the sense of a proposition [*Satz*] a thought and its meaning a truth-value. All true propositions have the same meaning: the true; and all false propositions have the same meaning: the false. (Frege, 1902/1983c, p. 149)

> [...] all propositions that express a true thought mean the same, and likewise all propositions that express a false thought. We have, e.g., $3 > 2. \supset .2^2 = 4$ and $2^2 = 4. \supset. 3 > 2$; consequently: $3 > 2. = .2^2 = 4.$ (Frege, 1903/1976, p. 158)

The task of the True and the False in Frege's logic is shown by the quotation from yet another letter: "Regarding the last points you touch on, I shall make the following: $\dot{\varepsilon}(—\varepsilon)$ is a class comprising only a single object, namely the true, and $\dot{\varepsilon}(\varepsilon = \neg\breve{\mathfrak{a}} \mathfrak{a} = \mathfrak{a})$ is a class comprising only a single object, namely the false" (Frege, 1902/1976b, p. 137).

Truth is essentially connected with the Law V, leading to antinomy. This law says: the equality of the value-ranges of two functions is equal to the general equality of those functions for every argument, in Frege's notation (1893/2009, p. 36):

$$\vdash(\dot{\varepsilon} f(\varepsilon)=\dot{\alpha} g(\alpha))=(\neg\breve{\mathfrak{a}} f(\mathfrak{a})=g(\mathfrak{a}))$$

Frege tried to save his system of logic against antinomy. In an afterword to the second volume of *Grundgesetze der Arithmetik*, he introduced a limitation of the generality of a function in defining the basic concepts of the arithmetic of natural numbers (Frege, 1903/2009c). The expressions the True and the False appeared there only once:

[14] For more on the Frege—Russell correspondence see (Besler, 2016).

> [...] the extension of a concept under which only the True falls should be the True and that the extension of a concept under which only the False falls should be the False. These determinations suffer no alteration under the new conception of the extension of a concept. (Frege, 1903/2009c, p. 263)

I would like to add that Frege was aware of a problematic aspect of the Law V ten years before Russell's discovery. He wrote:

> If anyone should believe that there is some fault, then he must be able to state precisely where, in his view, the error lies: with the basic laws, with the definitions, or with the rules or a specific application of them. If everything is considered to be in good order, one thereby knows precisely the grounds on which each individual theorem rests. As far as I can see, a dispute can arise only concerning my basic law of value-ranges (V), which perhaps has not yet been explicitly formulated by logicians although one thinks in accordance with it if, e.g., one speaks of an extension of a concept. I take it to be purely logical. At any rate, the place is hereby marked where there has to be a decision. (Frege, 1893/2009, p. VII)

It could be said that Frege doubted the truth of the Law V from the beginning (comp. Heck, 2010, p. 349–352), and unfortunately the Law V was crucial for his logistic program.

Geometry

Frege became acquainted with a new approach to geometry, which was David Hilbert's *Grundlagen der Geometrie* (1899). He was very impressed with this book, however, he could not agree with Hilbert. Frege did not accept (or did not understand) geometry understood as a formal system, allowing many models, including models of Euclidean geometry.

The topic of truth appears in the Frege—Hilbert correspondence in the context of different understanding of axioms in the system of geometry and their tasks. For Frege, axioms are true propositions. They do not need any proving, because "[...] our knowledge on them flows from a source very different from the logical source, a source which might be called spatial intuition. From the truth of the axioms it follows that they do not contradict one another" (Frege, 1899/1976, p. 37).

Frege assumed that all axioms of Euclidean geometry were irrefutable, he was convinced that "[...] it will be impossible to give such an example in the domain of elementary Euclidean geometry because all the axioms

are true in this domain" (Frege, 1900/1976, p. 71). Moreover, according to Frege, the axioms were necessarily consistent with each other. The truth and consistency of the axioms mutually conditioned each other.

Hilbert did not accept Frege's (idealistic) understanding of thought. Therefore, the philosophical background, always present in Frege's analysis, and rarely found in letters addressed to him, significantly disunited these correspondents. According to Frege, the thoughts-axioms are expressed in sentences-axioms (Blanchette, 2015, p. 111).

Similarly to many topics developed by Frege, the philosophy of language also appeared in the context of truth in geometry. In an unpublished paper on geometry from 1899–1906 he wrote:

> In the majority of cases what concerns us about thought is whether it is true [*Wahrsein*]. The most appropriate name for a true thought is a truth [*Wahrheit*]. A science is a system of truths [*Wahrheiten*]. A thought, once grasped, keeps pressing us for an answer to the question whether it is true [*Wahrsein*]. We declare our recognition of the truth of a thought, or as we may also say, our recognition of the truth [*Wahrheit*], by uttering a sentence with assertoric force. (Frege, 1899–1906/1983, p. 168)

After completing the correspondence with Hilbert, Frege returned several times to expressing his opinion on Hilbert's new approach to geometry. In one of the published articles, he repeated the thesis about the truth and consistency of axioms (Frege, 1903/1990).

It is worth pointing out at this juncture that Frege's comments on Hilbert's geometry were widely discussed, and Frege went down in the history of geometry as a defender of the truth of axioms (Freudenthal, 1957/2009, p. 494).

Expression generality

Frege combined truth with the expression of generality. In his logic, the generality of expressions is written in two ways:

1. Using the quantifier symbol.
2. By the appropriate type of variables.

As we may see, only the universal quantifier

$$,\smile_{\mathfrak{a}}\!\!- \Phi(\mathfrak{a})'$$

is introduced as a separate symbol. It was characterized in relation to truth-values as I have already written in this article.

Frege also used an existential quantifier (not calling it such), although it was absent as a separate sign. It was written with the use of the universal quantifier and negation-stroke, for example:

and read as "there is [*es gibt*] at least one square root of 1" (Frege, 1893/2009, p. 12). Probably due to the lack of a separate symbol for the existential quantifier, Frege connected truth only with the expression of generality and did not refer truth to the existential quantification.

Frege also signified generality by using an appropriate variable letter. For objects, they are letters of the Latin alphabet, *x*, *a*, etc. (Frege, 1893/2009, p. 11). He wrote:

> In order to obtain an expression for generality, one might have the idea of defining: "Let us understand '$\Phi(x)$' as the True if the value of the function $\Phi(\xi)$ is True for every argument; otherwise it shall refer to the False". (Frege, 1893/2009, p. 11)

For functions, they are capital letters of the Greek alphabet, Φ, Γ, etc. (Frege, 1893/2009, p. 35). There are also appropriate symbols for arguments of functions of higher degrees (Frege, 1893/2009, p. 60–61).

Truth is, for Frege, substantially connected with expressions of generality because, when assuming the above rules, they express the true thoughts (Frege, 1898–1903/1983, p. 162).

In this context, the assertion-stroke also appeared. Only sentences or formulas with a specific general domain can be preceded by the assertion-stroke, that is to say formulas with a quantifier or variables that express generality.

An example of a formula substantially connected with generality is the Law V, it expresses that equality of the value-ranges of two functions is equal to the general equality of those functions for every argument (Frege, 1893/2009, p. 36). After discovering that it leads to antinomy (Frege, 1902/1976a), Frege limited the domain of the functions, and in doing so he limited the scope of truth of the Law V (Frege, 1903/2009b, p. 262–263). At this juncture, as Frege saw it, he lost the generality of arithmetical propositions (Frege, 1903/2009b, p. 255).

Other Contexts

In the period examined here, the topic of truth also appears in other contexts, presented in Frege's correspondence with the great mathematicians of those times: David Hilbert (1862–1943) and Giuseppe Peano (1858–1932).

Frege's first letter to Hilbert concerned mathematical symbolism, and in this context the subject of truth appeared. Frege referred to mathematics understood as a game of symbols, in isolation from their references and wrote:

> A mere mechanical operation with formulas is dangerous (1) for the truth of the result and (2) for the fruitfulness of the science. The first danger can probably be avoided almost entirely by making the system of signs logically perfect. As far as the second danger is concerned, science would come to a standstill if the mechanism of formulas were to become so rampant as to stifle all thought. (Frege, 1895/1976, p. 33)[15]

Another thread comes from Frege's letter to Peano (undated, but surely written between 1896 and 1903) and concerns consequences resulting from various ways of defining equality in arithmetic, as a result of which

> [...] mathematicians agree indeed on the external form of their propositions but not on the thoughts they attach to them, and these are surely what is essential. What one mathematician proves is not the same as what another understands by the same sign. We only seem to have a large common store of mathematical truths [*Wahrheiten*], This is surely an intolerable situation, which must be ended as quickly as possible. (Frege, 1896–1903/1976, p. 126)

In these circumstances, Frege proposed, first of all, accepting identity, "complete coincidence" as the reference of the equality-sign (Frege, 1896–1903/1976, p. 126). Furthermore, thanks to distinguishing the equality at the level of sense from the equality at the level of reference, mathematics will be protected from generating always true, but boring instances of the principle of identity, $a = a$ (Frege, 1896–1903/1976, p. 126).

---

[15] In this letter Frege referred to his article (Frege, 1885/1900). It is not clear if Hilbert knew this paper.

Later Utterances on the Truth

Shortly after the publication of the position (1903b/2009) closing the period examined in this article, Frege introduced important new points to his theory of truth. It remains a separate topic as to how the problem of antinomy conditioned these changes, however, this topic requires a separate careful study.

Frege's penultimate letter to Russell is worth attention. There, Frege once again emphasised the particularity of the "true" predicate. What is more, there appears—for the first time—an excerpt, which can be treated as a basis for crediting Frege with a redundant understanding of truth:

> [...] "true" is not a predicate like "green". For at bottom, the proposition "It is true that 2+3=5" says no more than the proposition "2+3=5". Truth is not a component part of a thought, just as Mont Blanc with its snowfields is not itself a component part of the thought that Mont Blanc is more than 4000 high. (Frege, 1904/1976, p. 163)

Whether Frege actually assumed the redundant theory of truth and, if so, to what extent[16] is beyond the scope of this paper.

In the above-quoted letter there is also a clearly formulated principle of extensionality, referring to the substitutability *salva veritate* of expressions. This principle had already been used by Frege in his letter to Russell. He gave the following example of two propositions referring to the True: $2 + 3 = 5$ and $2 = 2$. Therefore it is correct to write: $(2 + 3 = 5) = (2 = 2)$ (Frege, 1893/2009, p. 9), where the sign "=" between the brackets shows the identity of the expression in brackets on the level of reference, but not on the level of sense. In the letter to Russell there is the following wording of the principle of extensionality used here:

> Then and only then does the meaning of the proposition enter into our considerations; it must therefore be most intimately connected with its truth. Indirect speech must here be disregarded. Disregarding it, we can therefore say that true proposition can be replaced by any true proposition without detriment to its truth, and likewise any false proposition by any false proposition. (Frege, 1904/1976, p. 165)

[16] For example Baldwin maintained that Frege did not assume the deflationary theory of truth (Baldwin, 1997, p. 9).

In 1918, Frege received from Ludwig Wittgenstein (1889–1951) a manuscript of *Logisch-philosophische Abhandlung* (Wittgenstein, 1921/1997) and inspired by it he published an article (Frege, 1918–1919/1990a). There he repeated many of his theses from the earlier paper (Frege, 1897/1983), adding arguments against the correspondence theory of truth[17] and the expression "the realm of thought" and its philosophical description.

Conclusion

For Frege, during the period of publishing the two volumes of *Grundgesetze der Arithmetik*, truth was an important category in the fields of philosophy of language (he starts from this aspect), formal logic (here truth plays the role of the key "tool"), philosophy of logic (expression of generality and the problem of antinomy), philosophy of mathematics (the problem of true axioms in geometry and understanding of equality in arithmetic) and ontology (idealistic understanding of the realm of thought, which is really connected with the truth). Bearing in mind the development of Frege's views on truth (Sluga, 2002), here I collect his main theses from the investigated period:

1. At the starting point, truth is examined on the basis of an ordinary language.
2. In logic, truth is expressed by the assertion-sign, therefore, truth is connected with judging.
3. Truth refers to logic more than any other science.
4. Truth is a normative category, because logical laws—as true—determine the direction of thinking.
5. Language expressions have sense and reference, the reference of sentences is truth-value (the True or the False).
6. Logic connectives, quantifier and logical entailment are characterized by truth-values.
7. The truth-bearers are: first thought, then proposition (or sentence) and language of science. Never ideas [*Vortsellungen*].

[17] Frege's criticism of the correspondence theory of truth was presented in (Sluga, 2007, p. 4–9; Baldwin, 1997).

8. The philosophical notion of thought is essentially connected with truth, as a certain unchanging objective ideal reality.
9. Truth is a primitive, indefinable term.
10. Truth—next to consistency—is an important notion describing geometrical axioms.

REFERENCES

Baldwin, T. (1997). Frege, Moore, Davidson: The Indefinability of Truth. *Philosophical Topics*, *25*(2), 1–18.

Besler, G. (2010). *Gottloba Fregego koncepcja analizy filozoficznej*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Besler, G. (2016). Philosophical and Mathematical Correspondence Between Gottlob Frege and Bertrand Russell in the Years 1902–1904. Some Uninvestigated Topics. *Folia Philosophica*, *35*, 85–100.

Blanchette, P. (2015). Frege's Critique of Modern Axioms. In: D. Schott (Ed.), *Frege: Freund(e) und Feind(e). Proceedings of the International Conference 2013* (pp. 105–120). Berlin: Logos Verlag.

Burge, T. (2005). *Truth, Thought, Reason. Essays on Frege*. Oxford: Clarendon Press.

Cook, R. (2016). Appendix: How to read *Grundgesetze*. In: G. Frege (2016), pp. A-1–A-42.

Dummett, M. (1981). *The Interpretation of Frege's Philosophy*. Cambridge: Harvard University Press.

Frege, G. (1879/1997). *Begriffsschrift und andere Aufsätze*. Hrsg. I, Angelelli. Hildesheim, Zürich, New York: Georg Olms Verlag. English edition: Conceptual Notation. In: G. Frege (1972), pp. 101–203.

Frege, G. (1885/1990). Über formale Theorien der Arithmetik. In: G. Frege (1990), pp. 103–111.

Frege, G. (1892/1990a). Über Begriff und Gegenstand. In: G. Frege (1990), pp. 167–178. English edition: On Concept and Object. In: G. Frege (1984), pp. 182–194.

Frege, G. (1892/1990b). Über Sinn und Bedeutung. In: G. Frege (1990), pp. 143–162. English edition: On Sense and Meaning. In: G. Frege (1984), pp. 157–177.

Frege, G. (1893/2009). Grundgesetze der Arithmetik. Begriffsschrift abgeleitet. Bd. 1. In: G. Frege (2009), pp. 1–303. English edition: G. Frege (2016), pp. I–XXXII, 1–253.

Frege, G. (1896/1990). Über die Begriffsschrift des Herrn Peano und meine eigene. In: G. Frege (1990), pp. 220–233. English edition: On Mr. Peano's Conceptual notation and My Own. In: Frege, G. (1984), pp. 234–248.

Frege, G. (1899/1990). Über die Zahlen des Herrn H. Schubert. In: G. Frege (1990), pp. 240–261. English edition: On Mr. H. Schubert's Numbers. In: G. Frege (1984), pp. 249–272.

Frege, G. (1903/2009a). Grundgesetze der Arithmetik. Begriffsschrift abgeleitet. Bd. 2. In: G. Frege (2009), pp. 305–583. English edition: Frege (2016), pp. I–XV, 1–266.

Frege, G. (1903/2009b). Nachwort. In: G. Frege (2009), pp. 549–563. English edition: Afterword. In: G. Frege (2016), pp. 253–265.

Frege, G. (1903/1990). Über die Grundlagen der Geometrie. In: G. Frege (1990), pp. 262–266. English edition: On the Foundation of Geometry. First Series. In: G. Frege (1984), pp. 273–284.

Frege, G. (1908/1990). Die Unmöglichkeit der Thomaeschen formalen Arithmetik aufs neue nachgewiesen. In*:* G. Frege (1990), pp. 329–333.

Frege, G. (1918–1919/1990a). Der Gedanke. Eine logische Untersuchung. In*:* G. Frege (1990) s. 342–361. English edition: Logical Investigation. I Thoughts. In: G. Frege (1984), pp. 351–372.

Frege, G. (1918–1919/1990b). Die Verneinung. Eine logische Untersuchungen. In*:* G. Frege (1990), pp. 362–378. English edition: Logical Investigation. II Negation. In: G. Frege (1984), pp. 373–389.

Frege, G. (1923/1990). Gedankengefüge. In: G. Frege (1990), pp. 378–394. English edition: Logical Investigation. III Compound Thoughts. In: G. Frege (1984), pp. 390–406.

Frege G. (1972). *Conceptual Notation and Related Articles*. Ed., transl., Bibliography, Introduction T. W. Bynum. Oxford: At the Clarendon Press.

Frege, G. (1976). *Wissenschaftlicher Briefwechsel*. Hrsg. G. Gabriel, H. Hermes, F. Kambartel, Ch. Thiel, A. Veraart. Hamburg: Felix Meiner Verlag.

Frege, G. (1895/1980). Frege to Hilbert, 1.10. In: G. Frege (1980), pp. 58–59. English edition: Frege to Hilbert. In: G. Frege (1980), pp. 32–34.

Frege, G. (1896/1976). Frege an Peano, 29.09. In: G. Frege (1976), pp. 181–186. English edition: Frege to Peano. In: G. Frege (1980), pp. 112–118.

Frege, G. (1896–1903/1976). Frege an Peano, n.d. In: G. Frege (1976), pp. 194–198. English edition: Frege to Peano. In: G. Frege (1980), pp. 125–129.

Frege, G. (1899/1976). Frege an Hilbert, 27.12. In: G. Frege (1976), pp. 60–64. English edition: Frege to Hilbert. In: G. Frege (1980), pp. 34–38.

Frege, G. (1900/1976). Frege an Hilbert, 6.01. In: G. Frege (1976), pp. 70–76. English edition: Frege to Hilbert. In: G. Frege (1980), pp. 43–48.

Frege, G. (1902/1976a). Frege an Russell, 22.06. In: G. Frege (1976), pp. 212–215. English edition: Frege to Russell. In: G. Frege (1980), pp. 131–133.

Frege, G. (1902/1976b). Frege an Russell, 29.06. In: G. Frege (1976), pp. 217–219. Frege to Russell. In: G. Frege (1980), pp. 135–137.

Frege, G. (1902/1976c). Frege an Russell, 20.10. In: G. Frege (1976), pp. 231–233. Frege to Russell. In: G. Frege (1980), pp. 149–150.

Frege, G. (1903/1976). Frege an Russell, 21.05. In: G. Frege (1976), pp. 239–241. Frege to Russell. In: G. Frege (1980), pp. 156–158.

Frege, G. (1904/1976). Frege an Russell, 3.11. In: G. Frege (1976), pp. 243–248. Frege to Russell. In: G. Frege (1980), pp. 160–166.

Frege G. (1979). *Posthumous Writing*. Eds. H. Hermes, F. Kambartel, F. Kaulbach. Trans. P. Long, R. White. Oxford: Basil Blackwell.

Frege G. (1980). *Philosophical and Mathematical Correspondence*. Eds. G. Gabriel, H. Hermes, F. Kambartel, Ch. Thiel, A. Veraart. Abridged for the English ed. B. Mcguinness. Transl. H. Kaal. Oxford, Basil Blackwell.

Frege, G. (1983). *Nachgelassene Schriften*. Hamburg: Felix Meiner Verlag.

Frege, G. (1879–1891/1983). Logik. In: G. Frege (1983), pp. 1–8. English edition: In: G. Frege (1979), pp. 1–8.

Frege, G. (1897/1983). Logik. In: G. Frege (1983), pp. 137–163. English edition: Logic. In: G. Frege (1979), pp. 126–151.

Frege, G. (1897–1898/1983). Begründung meiner strengeren Grundsätze des Definierens. In: G. Frege (1983), pp. 164–170.

Frege, G. (1898–1903/1983). Logische Mängel in der Mathematik. In: G. Frege (1983), pp. 171–181. English edition: Logical Defects in Mathematics. In: G. Frege (1979), pp. 157–166.

Frege, G. (1899–1906/1983). Über Euklidische Geometrie. In: G. Frege (1983), pp. 182–184. English edition: On Euclidean Geometry. In: G. Frege (1979), pp. 167–169.

Frege, G. (1906/1983). Einleitung in die Logik. In: G. Frege (1983), pp. 201–212. English edition: Introduction to Logic. In: G. Frege (1979), pp. 185–196.

Frege, G. (1914/1983). Logik in der Mathematik. In: G. Frege (1983), pp. 219–270. English edition: Logic in Mathematics. In: G. Frege (1979), pp. 203–250.

Frege, G. (1984). *Collected Papers on Mathematics, Logic, and Philosophy*. Ed. B. McGuinness. Oxford: Basil Blackwell.

Frege, G. (1990). *Kleine Schriften*. Hrsg. I. Angelelli. Hildesheim: Georg Olms.

Frege, G. (2009). *Grundgesetze der Arithmetik. Begriffsschrift abgeleitet. Bd. 1 und 2, in moderne Formelnotation transkribiert und mit einem ausführlichen Sachregister versehen von T. Müller, B. Schröder, R. Stuhlmann-Laeisz*. Paderborn: Mentis.

Frege, G. (2016). *Basic Laws of Arithmetic. Derived Using Concept-Script*. Eds., transl. Ph. Ebert, M. Rossberg, C. Wright. Oxford: Oxford University Press.

Freudenthal, H. (2009). Selecta. Retrieved from: https://www.maths.ed.ac.uk/~v1ranick/papers/freudselecta.pdf

Freudenthal, H. (1957/2009). Zur Geschichte der Grundlagen der Geometrie. Zugleich eine Besprechung der 8. Aufl. von Hilbert's "Grundlagen der Geometrie". In: H. Freudenthal (2009), pp. 486–523.

Glanzberg, M. (2018). *The Oxford Handbook of Truth*. Oxford: Oxford University Press.

Greimann, D. (2000). The Judgement-Stroke as a Truth-Operator. A New Interpretation of the Logical Form of Sentences in Frege's Scientific Language. *Erkenntnis*, *52*(2), 213–238.

Greimann, D. (2003a). *Das Wahre und das Falsche*. Hildesheim, Zűrich, New York: Georg Olms Verlag.

Greimann, D. (2003b). *Freges Konzeption der Wahrheit. Hildesheim*. Zűrich, New York: Georg Olms Verlag.

Greimann, D. (Ed.). (2007). *Essays on Frege's Conception of Truth*. Amsterdam, New York: Rodopi.

Heck, R. (2010). Frege and Semantics. In: M. Potter, T. Ricketts (Eds.), *The Cambridge companion to Frege* (pp. 342–378). New York: Cambridge University Press.

Heck, R. G., May, R. (2018). Truth in Frege. In: M. Glanzberg (Ed.), *The Oxford Handbook of Truth* (pp. 193–215). New York: Oxford University Press.

Hilbert, D. (1899). *Grundlagen der Geometrie*. Leipzig: Verlag von B. G. Teubner.

Marek, I. (1993). Początki matryc logicznych. *Logika*, *15*, 5–44.

Potter, M., Ricketts T. (Eds.). (2010). *The Cambridge Companion to Frege*. Cambridge: Cambridge University Press.

Reck E. H. (2002). *From Frege to Wittgenstein. Perspectives on Early Analytic Philosophy*. Oxford: Oxford University Press.

Russell, B. (1902/1976). Russell and Frege, 16.06. In: G. Frege (1976), pp. 211–212. English edition: Murawski (1986), pp. 221–222.

Schott D. (Ed.). (2015). *Frege: Freund(e) und Feind(e). Proceedings of the International Conference 2013*. Berlin: Logos Verlag, pp. 105–120.

Sluga H. (2002). Frege on Indefinability of Truth. In: E. H. Reck (Ed.), *From Frege to Wittgenstein. Perspectives on Early Analytic Philosophy* (pp. 75–95). Oxford: Oxford University Press.

Weiner J. (2002). Section 31 Revisited. Frege's Elucidations. In: E. H. Reck, *From Frege to Wittgenstein. Perspectives on Early Analytic Philosophy*
(pp. 149–182). Oxford: Oxford University Press.

Wittgenstein, L. (1921/1997). *Tractatus logico-philosophicus. Logisch-philosophische Abhandlung*. Leipzig: Unesma. English edition: *Tractatus logico-philosophicus*. New York: Dover Publications, Inc.

Originally published as "Gottlob Frege o prawdzie w okresie wydawania dwóch tomów *Grundgesetze der Arithmethik* (1893-1903)". *Studia Semiotyczne*, *32*(2), 51–73, DOI: 10.26333/sts.xxxii2.04. Translated by Gabriela Besler.

Studia Semiotyczne—English Supplement, t. XXX (2019), s. 85–106
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/stsen.xxx.05
Article

Krzysztof Wójtowicz*

THE NOTION OF EXPLANATION IN GÖDEL'S PHILOSOPHY OF MATHEMATICS[1]

Summary: The article deals with the question of in which sense the notion of explanation (which is rather characteristic of empirical sciences) can be applied to Kurt Gödel's philosophy of mathematics. Gödel, as a mathematical realist, claims that in mathematics we are dealing with facts that have an objective character (in particular, they are independent of our activities). One of these facts is the solvability of all well-formulated mathematical problems—and this fact requires a clarification. The assumptions on which Gödel's position is based are: (1) metaphysical realism: there is a mathematical universe, it is objective and independent of us; (2) epistemological optimism: we are equipped with sufficient cognitive power to gain insight into the universe. Gödel's concept of a solution to a mathematical problem is much broader than of a mathematical proof—it is rather about finding reliable axioms that lead to a (formal) solution of the problem. I analyse the problem presented in the article, taking as an example the continuum hypothesis.

Keywords: mathematical realism, mathematical explanation, incompleteness theorems, mathematical universe, continuum hypothesis.

* University of Warsaw, Institute of Philosophy. E-mail: wojtow@uw.edu.pl. ORCID: 0000-0002-1187-8762.

[1] The preparation of this paper (and its Polish version) was supported by National Science Centre (NCN) grant 2016/21/B/HS1/01955.

One of the theses put forward by Gödel is that about the solvability of all well-formulated mathematical problems. From the point of view of experience in the field of "everyday" mathematics (including school mathematics), this thesis seems obvious: every task can be solved, even very difficult open problems eventually give way to the pressure of the efforts of generations of mathematicians. However, Gödel is the author of the theorem that for each (reasonable) theory T, there are propositions that are undecidable in this theory. How can we reconcile this result with his thesis on the universal solvability of problems? In order to answer this question, a certain explication of the concept of solving a mathematical problem is necessary. Then it will be possible to analyse the thesis, according to which every problem would be solvable. How to explain it—and what explanation for this state of affairs is given by Gödel? I think that using the category of explanation here is justifiable. It is more and more often discussed in relation to mathematics—here it will have some specificity, but I think that its use will shed new light on the issue.

The article has the following structure:

1. Gödel's philosophy of mathematics.
2. The problem of explanation in mathematics.
3. The example of the continuum hypothesis.
4. Summary.

In part 1, I point to the basic elements of Gödel's philosophical worldview. The presentation is of course—necessarily—brief. In Part 2, I formulate the basic questions posed in the debate, I also briefly mention the problem of mathematical explanations in the natural sciences—and I formulate the title question/s. Part 3 is devoted to the analysis of the issue on the basis of a standard and well-known example—namely the continuum hypothesis. The article ends with a short summary.

1. Gödel's Philosophy of Mathematics[2]

Gödel was in a way, a model mathematical Platonist. In his opinion, there is an objective, mathematical universe independent of us, which is

[2] This is a very brief and sketchy presentation. A detailed analysis of Gödel's philosophical position is contained, for example, in the works of Krajewski (2003) and Wójtowicz (2002).

described (although, of course, in an imperfect way) through mathematical theories—and to which we have cognitive access through a kind of intuition.[3] Gödel focused on set theory, and his philosophical analyses often refer to it.[4] Gödel's views on the nature of mathematics naturally combine with a broader vision regarding the role and nature of philosophy. Gödel stressed the importance of fundamental analyses, in particular, analyses of the meaning of basic metaphysical concepts. He even hoped that he could describe these terms in an axiomatized way.[5] It is worth emphasizing his clear opposition to the dominant neo-positivist vision of mathematics (and philosophy, in particular metaphysics). Gödel even argued that the "spirit of the times" (*Zeitgeist*) is not in favour of his views that metaphysical considerations are meaningful and that mathematics is not the syntax of the language of science, but expresses objective truths. Conventionalism is not a good explanation for the nature of mathematics; conventions are, of course, present in mathematics, but they are not arbitrary, but—freely speaking—they convey the essence of concepts and express objective truths.[6]

---

[3] "Despite their remoteness from sense experience, we do have something like a perception also of the objects of set theory, as is seen from the fact that the axioms force themselves upon us as being true. I don't see any reason why we should have less confidence in this kind of perception, i.e., in mathematical intuition, than in sense perception, which induces us to build up physical theories and to expect that future sense perceptions will agree with them, and, moreover, to believe that a question not decidable now has meaning and may be decided in the future" (Gödel, 1964, pp. 120–121).

[4] Gödel's philosophical worldview was clearly reflected in his methodological decisions regarding how (by which methods) mathematics can be practised. Gödel declared that the belief in the existence of an objective mathematical world constituted the motivation for the free use of non-constructive methods based on strong assumptions about the existence of objects of a certain type.

[5] Gödel's proofs for the existence of God can be considered an attempt at this type of precision. Wang talks about the conversation between Gödel and Carnap on the 13th of September, 1940 (1987, p. 217), the subject of which was metaphysics, in particular the creation of a coherent metaphysical doctrine based on the notions of God and the soul as primitive. In Carnap's opinion, such a theory would have a mythological character, whereas Gödel's position is completely different. He claims that such a theory could be no less sensible than theoretical physics, which cannot be expressed in purely observational terms.

[6] The discussion of "syntactic interpretation" is devoted, for example, to the work in which Gödel writes: "in whatever manner the syntactic rules are formu-

Sometimes Gödel's position is presented as an expression of some kind of dogmatism—through a certain type of "act of faith" we postulate the existence of a mathematical universe to which mathematical propositions refer. Such a position would resemble the "working hypothesis" of many mathematicians—those who to the eternal question of whether mathematics is discovered or created, answer discovered (which is consistent with the position of realism, and can even be interpreted as one formulation of the realistic position). It would be an expression of a certain type of natural ontological position of a mathematician—but without any further justification.[7] However, Gödel did not accept this position in a dogmatic or non-reflective manner. It is worth noting a quite unusual—considering the conception of Gödel—and probably little-known quotation: "Our axioms, if interpreted as meaningful statements, necessarily presuppose a kind of Platonism, which cannot satisfy any critical mind" (Gödel, 1933, p. 50). We do not find such sceptical statements very much, but they document the fact that Gödel was aware that accepting a realistic position requires justification (and, of course, more precision—because realism can take many different forms). This may testify to a certain evolution of Gödel's views. He writes very clearly about this:

> Some body of unconditional mathematical truth must be ackhnowledged, because, even if mathematics is interpreted to be a hypothetical-deductive system, still the propositions which state that the axioms imply the theorems must be unconditionally true. The field of unconditional mathematical

---

lated, the power and usefulness of the mathematics resulting is proportional to the power of mathematical intuition necessary for their proof of admissibility. [...] it is clear that mathematical intuition cannot be replaced by conventions, but only by conventions plus mathematical intuition" (Gödel, 1953/9, p. 358).

[7] "However, when I do mathematics, I have a subjective feeling that there is a real world to discover: the world of mathematics. This world is much more imperishable for me, immutable and real than the facts of physical reality" (L. Bers, in: [Hammond, 1978, p. 19]). Hardy: "Personally, I always considered the mathematician in the first place as an observer, a man who observes a distant mountain range and notes his observations. His task is to clearly identify and describe to others as many peaks as possible" (Hardy, 1929, p. 18). Cantor talked about himself as a rapporteur for the results of his research. The conviction that the world of mathematical entities exists objectively—and we only discover it—connects all these mathematicians. Of course, I'm not saying that this position is the only one—or even that it is the dominant position among mathematicians, but that is a separate issue.

> truths is delimited very differently by different mathematicians. At least eight standpoints can be distinguished. [...]: (1) classical mathematics in the broad sense (i.e. set theory included), (2) classical mathematics in a strict sense, (3) semi-intuitionism, (4) intuitionism, (5) constructivism, (6) finitism, (7) restricted finitism, (8) implicationism. (Gödel, 1953/9, p. 346)

Gödel's argumentative strategy consists in adopting a weak version of realism as an initial assumption—and then a gradual strengthening of the position by indicating the relevant arguments. Number theory is a natural choice for this initial assumption because it is fundamental in mathematics—and widely known. Number theoretic propositions seem to express objective content.[8] The assumption that number-theoretic propositions have an objective character seems to be relatively uncontroversial. This is clearly stated in the following quote:

> Logic and mathematics—like physics—are based on axioms that have real content [...]. That such real content exists is evident through the study of number theory. We come across facts that are independent of any conventions. These facts must have content, because the consistency of number theory cannot be based on trivial facts. [...] There is a weak form of Platonism that no one can deny. [...] When we compare the Goldbach hypothesis with the continuum hypothesis, we are more convinced that the first of them must be true or false. (Gödel's statement in: Wang, 1996, pp. 211–212)

This opinion is significant in the context of Gödel's first and second theorems, according to which Peano arithmetic (PA) is imcomplete and its own consistency cannot be proved. Gödel's sentence (constructed in the proof) expresses—freely speaking—its own unprovability. We perceive it as true, but of course this is already due to a semantic analysis, going beyond the formal PA arithmetic. According to Gödel, such argumentation is fully legitimate (although it is not formalizable in PA). The source of mathematical knowledge is the analysis of concepts. It is based on the specific cognitive ability of our mind, i.e. mathematical intuition. This leads us to ever stronger theories, which we have the right to give realistic interpretations.

---

[8] It seems relatively natural to recognize that the truths of number theory have a "hard" character, that they are not just a matter of convention. The thesis that there exist n! permutations of the n-element set seems to be objective—and not the result of a purely conventional assumption.

2. Problems of Explanation in Mathematics

The problem of mathematical explanation is found in (at least) two areas: (i) mathematical explanations in natural sciences; (ii) explanations inside mathematics. Here, I focus exclusively on the issue of (ii). Probably the most natural version of this issue is the question about the explanatory nature of mathematical proofs: can (should?) mathematical proof play an explanatory role—and what does that mean? It is clear that the basic function of proofs is to convince (in accordance with the standards of mathematical argumentation) that a theorem is true. At the same time, the natural (but not strictly formal) question for a mathematician is one about deeper causes, about the whole "background of phenomena". Speaking freely, when analysing mathematical proofs, it is important not only how the individual inferential steps follow from each other, but "what's really going on here?". Using somewhat metaphorical language, it is about this subtle "game of mathematical concepts", which does not boil down to the fact that the next step of the proof results from the previous one. Understanding mathematical proof as a formal verification of facts (by examining formal dependencies) does not fully reflect the understanding of mathematical proof as a source of mathematical knowledge. Sometimes mathematicians speak in such a spirit:

> Even when a proof has been mastered, there may be a feeling of dissatisfaction with it, though it may be strictly logical and convincing; [...]. The reader may feel that something is missing. The argument may have been presented in such a way as to throw no light on the why and wherefore of the procedure or on the origin of the proof or why it succeeds. (Mordell, 1959, p. 11; citation based on: Mancosu, 2008, p. 142)

Similarly, Rota writes (in the context of computer evidence) that "[v]erification is proof, but verification may not give the reason" (Rota, 1997, p. 187).[9] The question about the explanatory role of mathematical proofs has a long history—as early as in Aristotle one can find a distinction corresponding, in today's terminology, to reasoning that only justifies

[9] There is no room for detailed analysis of the issue. I consider Rav's article (1999), in which the author analyses the role of proofs in mathematics, accentuating its central place, to be very interesting.

a certain thesis and reasoning that explains the reasons.[10] Mathematicians themselves are obviously aware of the different nature and function of proofs. Mancosu (2018) gives an example of a monograph on algebraic geometry, which deals with various proof methods, and in which the author rejects the so-called the transfer method (despite its effectiveness), indicating that it allows to give a logical proof of a certain result, but does not explain it.[11] The discussion about explanations in mathematics is lively—there are many detailed analyses regarding individual theorems, the links between the problem of explanation and the (quite elusive but important) concept of depth in mathematics,[12] aesthetic issues, or the problem of purity of proofs (i.e. using methods limited to a given field—e.g. purely geometric methods in proofs of geometry theorems or combinatorial methods in combinatorial proofs). However, there is still no good general answer to the question of what is the real source of explanatory power of mathematical proofs.

The problem of explanation may also have a broader character—and may relate not only to the proofs, but even to broader classes of issues. The question "why is squaring the circle impossible?" has a slightly broader dimension: the answer can be found outside of geometry, in Galois's theory. Therefore, it is no longer a question of the proof only, but also or giving a proper interpretation of one theory in another. Similarly, you can ask questions about the nature of concepts which are fundamental for a given theory, about the most natural formulations (definitions), etc. This is a very broad issue and will not be addressed here.

This problem of explanation (or maybe: a series of problems) concerns explanations inside mathematics. However, the subject of analysis in this article is a question that is not mathematical *par excellence*—rather philosophical or methodological. The general question about why every mathematical problem is solvable has a completely different character to the very specific question, for example, why every differentiable function is continuous, or why squaring a circle or triscecting an angle is not possible.

---

[10] See, for example, Mancosu (2018), where the reader will find a detailed description of the problem of mathematical explanations (both in physics and in mathematics itself) together with a comprehensive and up-to-date bibliography. I thank one of the reviewers for drawing my attention to this.

[11] This monograph is Brumfiel (1979). In another work (Hafner & Mancosu, 2008), the authors analyse this example in the context of Kitcher's explanation theory.

[12] See special issue *23*(2) *Philosophia Mathematica* (2015).

In these cases, we primarily ask about the proof, or possibly its analysis and commentary (explanation): what "resources" we use, what assumptions are necessary (and what role do they play in the proof), which set of concepts we refer to, what is the "conceptual environment"? Ultimately, therefore, often the answer can be reduced to analysing some specific proof. On the other hand, it is difficult to expect a similar analysis of a philosophical thesis—especially in the context of the fact that the Gödel's first theorem seems to contradict this thesis at first glance.

However, in the context of Gödel's philosophy of mathematics, I consider using the concept of explanation in this context to be legitimate. The concept of solving a mathematical problem—according to Gödel exceeds the notion of formal proof. It should be remembered that Gödel considered technical and philosophical issues to be intimately connected.[13] It is worth recalling that Gödel believed that philosophical considerations could be given a clear form and that (after sufficiently good clarification of the concepts) philosophical discussion reaches the level of precision which is typical for mathematics (Gödel, 1951, p. 322). In such an optimistic spirit, one can interpret his statement that the design of Leibniz's *characteristica universalis* was not a pure utopia (Gödel, 1944, p. 101). At the same time, he admitted that this is a matter for the future and that, for the time being, philosophy has not reached a sufficient degree of development (Gödel, 1951, p. 311). He himself admitted that he did not give his analyses a sufficiently precise form.

We talk about explanation in a natural way when we are dealing with a phenomenon that we want to describe, understand or just explain. Usually (and certainly often) this phenomenon is something external, it is not a convention, for example physical phenomena are given to us, we are confronted with them. Will a similar approach be appropriate for mathematics, which seems to be our creation, though? In the context of Gödel's realistic position, such an approach is natural: mathematics is somewhat independent of us, it has an objective character. So it is not surprising that we are confronted with objective facts—also concerning mathematics. We want to explain these facts. An example of such a fact is the solvability of problems. Answering the question: "Why is every well-formulated mathematical problem solvable?" is associated with the need to clarify

[13] The creator of set theory, Cantor, argued that mathematical and philosophical problems cannot be separated—and that set theory would give a theological interpretation (e.g., Murawski, 1984; Purkert, 1989).

how to understand the concept of solvability (solution) of a mathematical problem. This issue can be "invalidated" by reducing it to a king of tautological statement: the problem is well-formulated exactly when it is solvable (even if we do not know this solution, or even—potentially—we will never know it). And here the discussion ends. However, I believe that would not be the right attitude to the matter. The concepts of "well-formulated problem" and "solution to the problem" are not easily reducible to each other—the history of mathematics shows clearly that it would be an over-simplification.

The concept of solving a mathematical problem from the point of view of ordinary, everyday mathematics has obvious meaning: to "solve the problem" is simply to provide the appropriate proof, using standard means. Probably for 99.9% of problems encountered by a mathematician in practice, this is what is meant by a solution. However, the situation becomes more complicated when we reach problems which are undecidable within standard mathematics. The question arises what standard mathematics is. The view that standard mathematics can be reconstructed in ZFC set theory (i.e. Zermelo-Fraenkel set theory with the axiom of choice)—and it is the ZFC that sets the framework of the "mathematical standard"—is quite common in the philosophy and foundations of mathematics. This point of view is very clearly visible in Gödel himself.

It has been known from the moment of proving Gödel's first theorem that ZFC is an incomplete theory, and the first example of an independent proposition with a clear mathematical content is the continuum hypothesis.[14] It is obvious, therefore, that the concept of solving a mathematical problem must have a different meaning to "deciding it within ZFC"—otherwise Gödel's thesis would be clearly and obviously false.

Gödel's position is worth considering in the context of Hilbert's programme and Hilbert's mathematical worldview. Hilbert was undoubtedly a cognitive optimist—he argued that there is no *ignorabimus* in mathematics and that any well-formulated mathematical problem can be

[14] Gödel's theorems talk about the existence of independent propositions, but the construction of Gödel's sentence does not lead to propositions with a natural mathematical content. CH is such a natural sentence which is independent of ZFC—and this is a very important result. It is worth adding that the first independent propositions from PA with a clear combinatorial content were given only in the 1970s (Paris & Harrington, 1977).

solved.[15] Hilbert's programme can also be seen as an expression of this optimism: he hoped to find a safe foundation for mathematics—which would also be strong enough to solve (all) well-formulated problems. Tools for this are to be provided by proof theory . Hilbert was, therefore, convinced that any mathematical problem could be solved in a literal sense (probably closest to the colloquial meaning).[16]

A common assertion in the literature is that Gödel's theorems dealt a fatal blow to Hilbert's programme. This is a suggestive statement, but probably Gödel would not agree with it himself, in any case not entirely. In his unpublished notes, he notes that interpreting finitist mathematics as a purely formal system leads to a dilemma (Gödel, 193?, p. 164). We can, therefore, say:

(i) that not every mathematical problem is solvable;

(ii) that the syntactic approach to proof does not constitute a proper representation of our concept of proof as something that is the source of our certainty and allows the solving of mathematical problems.

---

[15] The French physiologist, Emil du Bois-Reymond, in 1872, formulated the thesis of *ignorabimus*, according to which science is burdened with internal limitations, and so there must be problems impossible to solve. His brother was Paul du Bois-Reymond (an eminent mathematician) who considered this thesis also justified in relation to mathematics (McCarty, 2004). This brings Kant's attention to the questions agonising people's minds, which "one cannot suppress, because he is asked it by his own nature, but which he cannot answer because they outweigh all his potency" (Kant, 1957, p. 7).

[16] Slightly simplifying, it can be said that up to the turn of the 19th and 20th centuries there was no concept of formal proof, and mathematical proofs had—speaking freely—a semantic character. Only with the development of formal logic was it possible to formulate the concept of "formal proof" as a specific set of operations with a formal character (although beliefs of this type—in a yet undefined form—were already present in mathematics). A paradigmatic example, which very clearly shows the discrepancy between the traditional (semantic) and formal concept of proof, is geometry, which was formalized by Hilbert in *Grundlagen der Geometrie*. The formalistic point of view on geometric proofs obviously assumes that there is some established formal system in which these proofs are reconstructed and that this system encompasses all truths (or "truths"). There is no room for intuitive argumentation—for example Hahn was very radical against the concept of intuition.

Gödel points to the fact that

> number-theoretic questions which are undecidable in a given formalism are always decidable by evident inferences not expressible in the given formalism. As to the evidence of these new inferences, they turn out to be exactly as evident as those of the given formalism. So the result is rather that it is not possible to formalise mathematical evidence even in the domain of number theory, but the conviction about which Hilbert speaks remains entirely untouched (Gödel, 193?, p. 164)

thus advocating the second possibility. It can be said that, in his opinion, the syntactic interpretation leads to the loss of important aspects of the proof.

And just seeing this fact allows Gödel to remain a cognitive optimist with regard to mathematics. However, he interpreted the concept of "solution to a mathematical problem" in a radically different way from Hilbert. According to Gödel, convincing mathematical reasoning can be informal.[17] An example is the proposition constructed in the proof of Gödel's theorem: there is no doubt that the proposition "I am unprovable within PA" is perceived as true, although of course it is not provable within PA.

So, the notion of "resolving a mathematical problem" will be interpreted by Gödel in a very different way from Hilbert. It can be said that they interpret the term "mathematical knowledge" in a different way, or that they respond in a different way to the question "what does it mean to have mathematical knowledge?" From the point of view of the Hilbert programme, obtaining mathematical knowledge is possible thanks to the establishment of an unquestionable, finitary fragment of mathematics (and then by performing the appropriate theoretical reduction). For Gödel, the matter looks completely different—which is of course related to the incompleteness theorems. No formal theory (satisfying the relevant natural conditions) is a complete theory, and thus it will not be possible to solve all mathematical problems in one theory.) The process of obtain-

[17] It is worth mentioning again that, according to Gödel, it will be possible to conduct a philosophical discussion with mathematical accuracy (the condition is a good explanation of concepts; Gödel, 1951, p. 322). Wang cites Gödel's opinion that a precise metaphysical doctrine will be formulated in the future. Its absence results from the erroneous way of practising philosophy (and theology) as well as the prevailing scientistic superstitions (Wang, 1987, p. 159).

ing mathematical knowledge goes beyond formal procedures, and mathematical argumentation is not reducible to the concept of "proof in theory T". The proofs that we know from mathematical practice, of course, are not formal in nature: rather, they consist of convincing arguments in which an intuitive understanding of mathematical concepts is inevitably present—not only formal transformations. A spectacular example is the proof of Fermat's theorem—it is hard to imagine what it would look like in a fully formalized version, but it certainly would not be readable for us.[18]

The central concept in Gödel's philosophy of mathematics is mathematical intuition—a kind of intellectual ability to recognize mathematical truths, that goes beyond the mechanical manipulation of symbols. In this context, it is worth mentioning the important work of Turing (1939). Turing draws attention to the fact (in the context of Gödel's results) that we are able to see the truth of unprovable statements in a given formalism. In his work, he analyses the problem of the whole system of increasingly stronger logics, in which it will be possible to solve ever-wider classes of mathematical problems—which can also be understood as a technical equivalent of Gödel's idea going beyond the given formal system.[19] Regardless of how we are going to understand the concept of mathematical intuition, there is no doubt that it cannot be mechanical—and thus cannot be "imitated" in the standard model of the Turing machine. However, it can be argued (e.g., Hodges, 2013) that the concept of the oracle, introduced by Turing, is the formal equivalent of cognitive activities that go beyond mechanical procedures. Turing does not analyse the nature of the oracle in more detail, limiting himself to the statement that it cannot be a machine. It can, therefore, be said that the informal, intuitive component of the activity of the mathematician has been "incorporated" into the technical definition here.

There is a tension here between what we would call a "mathematically convincing argument" and its formal paraphrase (or perhaps: its *explica-*

[18] An interesting example of a proof that is short, understandable and fully acceptable is given by Boolos (1987). This is a proof in second order logic—but the formalization of this proof in first order logic would be "astronomical" in length. The problem of formalizing this proof in Mizar is the subject of analyses in the work of Benzmüller and Brown (2007). I thank one of the reviewers for drawing my attention to this issue and for the bibliographic suggestions—as well as for suggestions regarding Turing's work.

[19] In Marciszewski's essay (2018) this issue is discussed more comprehensively.

tum in the form of the concept of formal proof). The formalistic position (in the wide sense) reduces the notion of a mathematically correct argument to the notion of a formal proof in the relevant theory T. However, Gödel's position is completely different—from his point of view, well-formulated mathematical problems are not problems that are solvable within some specific theory T. Rather—freely speaking—for each well-formulated mathematical problem one can formulate the relevant theory T that will solve it. And, of course, it is not a trivial claim that if we have a proposition φ independent of the theory T, then within the theory $T + \varphi$ (i.e., T with φ added as a premise), this problem will be settled. The point is, of course, that it is possible to search for natural, mathematically justified theories T^*, being extensions of T—and resolving our (previously) undecidable propositions.

It is worth mentioning the discussion between Gödel and Zermelo regarding, *inter alia*, the issue of solving mathematical problems.[20] In a letter to Gödel of 21st September, 1931, Zermelo opposes the thesis that any mathematical notion can be defined by means of a finite series of symbols—he calls this conviction a "finitist prejudice". He even claims that Gödel's results express an obvious fact: if only countably many sentences can be defined in a formal language, and there are uncountably many truths, then obviously there must be unprovable truths. It can be argued that Zermelo underestimated the importance of Gödel's results and did not fully understand the technical subtleties. Gödel responds to Zermelo's letter (in a letter dated 12.10.1931), explaining what the essence of his proof consists of—and in particular, emphasizing that what is relevant are statements expressible in a given system, but unprovable in this system, and at the same time provable in a more powerful system. Zermelo interprets the use of a stronger system as a modification of the concept of proof itself. He argues that providing proof involves making the proved sentence obvious, which is achieved by formulating a suitable set of propositions. Zermelo poses a question about what this obviousness is—and at the same time formulates the hypothesis that in a suitable system every mathematical problem is solvable (letter to Gödel from 29.10.1931). The correspondence did not go any further, however, it is an interesting testimony to the early reception of Gödel's results. Another

[20] I thank one of the reviewers for drawing my attention to this issue, and for pointing out the work of Ebbinghaus, Fraser, Kanamori (2010), in which (on pages 482–501) the correspondence cited is included.

interesting point is the issue of problem solving: Gödel, being aware of the existence of metamathematical constraints, believes that it will be possible to establish new axioms that allow for the resolution of subsequent problems. On the other hand, according to Zermelo, these limitations are an obvious defect of the finitist systems, and mathematical reasoning should be reproduced in infinitary systems. This is in accordance with his well-known statement that the proper logic for mathematics is infinitary logic.[21]

3. The Example of the Continuum Hypotheses

Gödel distinguished between objective mathematics (as a set of truths about the mathematic universe) and subjective mathematics (i.e., that which is known to us). His realistic position assumed that the task of the mathematician is to search for a description of mathematical reality—which is objective and exists independently of us. Formal systems describe it only partially—and of course we cannot stop at one particular system as the final set of truths. Rather, it is necessary to analyze mathematical concepts (in particular—the concept of a set) so as to be able to justify new axioms—which will allow for the resolution of subsequent open problems. However, in the case of arithmetic itself, informal reasoning convinces us of the truth of, e.g., Gödel's proposition "I have no proof", while mathematical practice and our beliefs about arithmetic lead to the acceptance of Con(PA). But it would be difficult to give that type of natural and obvious intuitive argumentation in the case of propositions independent of set theory.

In search of an explanation of the solvability of any well-defined mathematical problem it is good to refer to a specific example—and in this article it will be the continuum hypothesis (CH), which is a paradigmatic example of a sentence independent of ZFC.[22] ZFC imposes few limitations: there are many propositions of the type "the value of the

[21] A very interesting description of Zermelo's infinitary logic programme can be found in Pogonowski's work (2006).

[22] The continuum hypothesis is that the power of the set of real numbers (i.e. the power of a continuum) is the smallest uncountable cardinal number, i.e. $\aleph_1$. In another formulation: each infinite subset of R is either countable or equinumerous with R. The independence of CH from ZFC was proven by Gödel and Cohen: Gödel showed its consistency with the ZFC axioms, and Cohen in 1963 the consistency of its negation.

continuum is $\aleph\alpha$" that are consistent with ZFC.[23] However, despite formal independence, one can ask whether there are any convincing arguments that would allow to assign a particular value to the continuum—and above all, whether the continuum problem is a well-posed mathematical problem.

In one of his best-known articles, Gödel analyses the continuum hypothesis (Gödel, 1964). He regards it as an objective, well-formulated question about mathematical reality.[24] It is obviously unprovable in ZFC, but this simply results from the weakness of this theory. For objective mathematics—i.e. all unconditionally true propositions—is one thing, and subjective mathematics: all probative propositions in a given formal theory, is another (Gödel, 1951, p. 305). He himself leant towards the thesis of the falsity of CH, pointing to its paradoxical consequences (Gödel, 1964). However, his views on this matter are not widely accepted. Gödel was, therefore, convinced that it would be possible to find axioms which will determine the value of the continuum. As it is known, the axiom of the constructability V = L implies CH (and also the generalized continuum hypothesis). V = L might be viewed as minimalistic (the universe of collections is "narrow"). So Gödel assumed that it would be possible to prove CH from some axiom of a maximalist character, in a sense opposite to V = L (Gödel, 1964, p. 266). In a certain well-defined sense, large cardinals axioms can be considered to be such maximalist axioms—and here Gödel hoped to find a solution. He was aware that strong axioms of this type would be needed, and that Mahlo numbers relatively low in the infinity hierarchy would not be sufficient.[25]

---

[23] There is a well-known theorem that shows how "strangely" the power of cardinal numbers can behave. Easton showed that for any F function meeting two conditions: (1) F is a non-decreasing function from the class of regular cardinal numbers in cardinal numbers; (2) for any κ: $\kappa < \mathrm{cf}(F(\kappa))$; a model for set theory can be constructed in which for any regular cardinal number κ, $2^{\kappa} = F(\kappa)$ (Easton, 1970). In particular, the continuum (that is 2^{ω}) can be large.

[24] Arguments in favour of the thesis that the continuum hypothesis is a well-formulated mathematical problem, not just a metamatematical one, are formulated, for example, by Hauser (2002).

[25] Gödel's article (1964) is not the only (or the first) place where he expressed such opinions. In a lecture at Princeton in 1946 Gödel characterized "strong infinity axioms" as an assumption which, in addition to having a specific formal structure, is "is also true" (Gödel, 1946, p. 151). He also expressed a very optimistic conjecture that "some completeness theorem would hold which would say that

It turned out that this this strategy would not bring success in solving the continuum problem: the results, according to which various strong large cardinal axioms are consistent with both the continuum hypothesis and its negation, are known (Levy & Solovay, 1967). Let us add here that Gödel himself tried to formulate another type of axiom that would solve this problem (Gödel, 1970a; 1970b).[26]

However, regardless of the fact that studies concerning large cardinals did not solve the continuum problem, the very idea of seeking new axioms became an inspiration to researchers, and, the Gödel programme is often referred to in this context. Of course—such axioms could not be *ad hoc*, but they would result from analyses regarding our understanding of the concept of the set and our vision of the mathematical universe. The discussion on this subject is lively—however, even a brief review definitely goes beyond the scope of this article.[27]

So when it comes to the *explicatum* defined above ("solvability of a mathematical problem"), one can be tempted to characterize it as finding the appropriate formal theory T—which is an extension of ZFC—based on natural, acceptable axioms, leading to the formal settlement of the problem P within T. There would be two components here:

- Conceptual-analytical phase: the search for appropriate natural, acceptable axioms—and the formulation of the relevant theory T.
- Technical phase: the resolution of P within T (i.e., standard mathematical work—perhaps very difficult).[28]

---

every proposition expressible in set theory is decidable from the present axioms plus some true assertion about the largeness of the universe of all sets" (Gödel, 1946, p. 151).

[26] According to commentators, Gödel's reasoning was mistaken (cf. Ellentuck, 1975; Solovay, 1995).

[27] We may mention, for example: Feferman, (1996; 2000), Friedman (2000), Maddy (1988a; 1988b; 1993; 1997), Steel (2000). Woodin's works (1999, 2001) contain technically very complex methodological analyses, based on which it can be proven that the continuum value is $\aleph_2$. Of course, they are the subject of discussion and controversy, so it cannot be argued that the continuum problem has been solved.

[28] Regarding the continuum hypothesis, he stated: "When the concept of set becomes clear, even when we find satisfactory infinity axioms, there will still be a technical (i.e. mathematical) problem to resolve the continuum hypothesis based on axioms" (Wang, 1996, p. 237).

What is the philosophical background for the belief that this is always possible and that every well-defined problem is solvable? Two important aspects can be identified here. One would be termed metaphysical, and the other methodological. Speaking of the metaphysical aspect, I mean that Gödel's realistic position presupposes the existence of an objective, mathematical universe with a certain nature. Gödel believed that the universe has a set-theoretic character—and that there is one, objective universe in which all mathematical propositions are interpreted, and moreover every proposition is either true or false in it. There are, therefore, no propositions of undetermined logical status, no "shaky" propositions.[29] Gödel's thesis would, therefore, have a metaphysical foundation in a specific vision of the mathematical universe.[30]

Of course, belief in the existence of one objective (though unknown) mathematical universe does not automatically give any clues as to what are the solutions to open mathematical problems. After all, it would be possible to accept the thesis that the mathematical world has an objective and fixed character, but that it is unknowable (that is, the *ignorabimus* thesis would be true, against the optimism of Hilbert or Gödel). And here we touch on the methodological aspect: the way in which we can seek answers to mathematical questions that are *ex definitione* unsolvable within the available, i.e. accepted, standard theory (e.g. ZFC). This is possible by establishing new, credible axioms. Gödel was convinced that our analysis of the concept of set would allow the establishment of such axioms. This is an expression of a specific epistemological vision: accord-

[29] It would be possible to think this if one adopted the concept of so-called multiverses—i.e. a realistic concept, according to which mathematical reality exists, but it is not a "uniform" mathematical universe, rather the entire "galaxy" of set theoretic universes that implement different concepts of set (e.g. Hamkins, 2012). In such a situation, it would not make sense to say that e.g., the continuum hypothesis has a logical value: in different universes the continuum could take different values.

[30] This article is not of a historical-exegetical nature, but it is worth noting that it seems that Gödel's opinion has undergone some evolution of view. He writes that "it is very plausible that with [V = L one is dealing with an absolutely undecidable proposition, on which set theory bifurcates into two different systems, similar to Euclidean and non-Euclidean geometry" (Gödel, 1939b, p. 155). Thus, he explicitly allows for the existence of absolutely insoluble problems; similar theses can be found in another text (Gödel, 193?). Undoubtedly, he later claimed that V = L should be rejected.

ing to Gödel, we have the ability to analyse concepts and see these truths. He regarded the phenomenological method as promising, and wrote about it explicitly in one of his works (Gödel, 1961; cf. also e.g. Tieszen, 1998).[31]

4. Conclusions

Gödel understands the concept of the solution of mathematical problems much more broadly than as the providing of mathematical proof. Formulating such a proof is obviously a necessary condition (and in the case of the vast majority of standard mathematical problems—sufficient), but there are also mathematical problems for which the formulation of a proof is only the second stage. The first is to find reliable (true!) assumptions on the basis of which this proof can be carried out. Obviously, these assumptions must go beyond the standard set theory, i.e. ZFC.

What though is the explanation for this phenomenon of problem solving? The first assumption on which Gödel's view is based is metaphysical realism: there is a mathematical universe, it is objective, independent of us—and each mathematical proposition has a logical value. The second assumption is a kind of epistemological optimism: we are equipped with sufficiently good cognitive means to gain insight into this universe.

The use of the notion of explanation, which is characteristic of empirical sciences, is justified: in the objectivistic vision of Gödel, we are dealing with facts that are independent of us. One of these facts is the solvability

[31] It is worth mentioning here the "second pillar" of learning mathematical truths—they can be methodological arguments that can be symbolically labelled "fruitfulness". This is a very broad issue that I shall not analyse here. It is worth remembering that Gödel himself very clearly emphasized the importance of this aspect, as evidenced by the following quote: "a probable decision about its [a new axiom—K.W.] truth is possible also in another way, namely, inductively by studying its 'success'. Success here means fruitfulness in consequences, in particular in 'verifiable' consequences, i.e., consequences demonstrable without the new axiom, whose proofs with the help of the new axiom, however, are considerably simpler and easier to discover, and make it possible to contract into one proof many different proofs. [...] There might exist axioms so abundant in their verifiable consequences, shedding so much light upon a whole field, and yielding such powerful methods for solving problems (and even solving them constructively, as far as that is possible) that, no matter whether or not they are intrinsically necessary, they would have to be accepted at least in the same sense as any well established physical theory" (Gödel, 1964, pp. 113–114).

of all well-formulated mathematical problems—and this fact requires explanation.

REFERENCES

Brumfiel, G. W. (1979). *Partially Ordered Rings and Semi-Algebraic Geometry*. Cambridge: Cambridge University Press.

Easton, W. B. (1970). Powers of Regular Cardinals. *Annals of Mathematical Logic*, *1*(2), 139–178.

Ebbinghaus, H-D., Fraser, C. G., Kanamori, A. (2010). *Ernst Zermelo. Collected Works. Gesammelte Werke. Vol. I.* Berlin, Heidelberg: Springer-Verlag.

Ellentuck, E. (1975). Gödel's Square Axioms for the Continuum. *Mathematische Annalen*, *216*(1), 29–33.

Feferman, S. (2000). Why the Programs for New Axioms Need to Be Questioned. *The Bulletin of Symbolic Logic*, *6*, 401–413.

Friedman, H. (2000). Normal Mathematics Will Need New Axioms. *The Bulletin of Symbolic Logic*, *6*(4), 434–446.

Gödel, K. (193?). Undecidable Diophantine Propositions. In: S. Feferman (Ed.), *Kurt Gödel: Collected Works: Volume III* (pp. 164–175). Oxford: Oxford University Press.

Gödel, K. (1933). The Present Situation in the Foundations of Mathematics. In: S. Feferman (Ed.), *Kurt Gödel: Collected Works: Volume III* (pp. 45–53). Oxford: Oxford University Press.

Gödel, K. (1939b). Vortrag Göttingen (Lecture at Göttingen). In: S. Feferman (Ed.), *Kurt Gödel: Collected Works: Volume III* (pp. 127–155). Oxford: Oxford University Press.

Gödel, K. (1944). Russell's Mathematical Logic. In: P. A. Schlipp (Ed.), *The philosophy of Bertrand Russell. Library of Living Philosophers* (vol. 5, pp. 123–153). La Salle, Illinois: Open Court Publishing Company.

Gödel, K. (1946). Remarks Before the Princeton Bicentennial Conference on Problems in Mathematics. In: S. Feferman (Ed.), *Kurt Gödel: Collected Works: Volume II* (pp. 150–153). Oxford: Oxford University Press.

Gödel, K. (1951). Some Basic Theorems on the Foundations of Mathematics and Their Implications. In: S. Feferman (Ed.), *Kurt Gödel:*

Collected Works: Volume III (pp. 304–323). Oxford: Oxford University Press.

Gödel, K. (1953/9). Is Mathematics Syntax of Language? In: S. Feferman (Ed.), *Kurt Gödel: Collected Works: Volume III* (pp. 334–363). Oxford: Oxford University Press.

Gödel, K. (1961). The Modern Development of the Foundations of Mathematics in the Light of Philosophy. In: S. Feferman (Ed.), *Kurt Gödel: Collected Works: Volume III* (pp. 374–387). Oxford: Oxford University Press.

Gödel, K. (1964). What is Cantor's Continuum Problem? In: P. Benacerraf, H. Putnam (Eds.), *Philosophy of Mathematics: Selected Readings* (pp. 258–272). Englewood Cliffs, New Jersey, Prentice-Hall, Inc.

Gödel, K. (1970a). Some Considerations Leading to the Probable Conclusion, That the True Power of the Continuum Is $\aleph_2$. In: S. Feferman (Ed.), *Kurt Gödel: Collected Works: Volume III* (pp. 420–421). Oxford: Oxford University Press.

Gödel, K. (1970b). A Proof of Cantor's Continuum Hypothesis from a Highly Plausible Axiom About Orders of Growth. In: S. Feferman (Ed.), *Kurt Gödel: Collected Works: Volume III* (pp. 422–423). Oxford: Oxford University Press.

Hafner, J., Mancosu, P. (2008). Beyond Unification, In: P. Mancosu (Ed.), *Philosophy of Mathematical Practice* (pp. 151–178). Oxford: Oxford University Press.

Hamkins, J. D. (2012). The Set-Theoretic Multiverse. *Review of Symbolic Logic*, *5*(3), 416–449.

Hammond A. L. (1978). Mathematics—Our Invisible Culture. In: L. Steen (Ed), *Mathematics Today. Twelve Informal Essays* (pp. 15–34.). New York, Heidelberg, Berlin: Springer-Verlag.

Hardy, G. H. (1929). Mathematical Proof. *Mind*, *38*(149), 1–25.

Hauser, K. (2002). Is Cantor's Continuum Problem Inherently Vague? *Philosophia Mathematica*, *10*(3), 257–292.

Hodges, W. (2013). Alan Turing. *The Stanford Encyclopedia of Philosophy*. Retrieved from: https://plato.stanford.edu/archives/win2013/entries/turing/

Kant, I. (1957). *Krytyka czystego rozumu* (vol. 1). Warszawa: PWN.

Krajewski, S. (2003). *Twierdzenie Gödla i jego interpretacje filozoficzne*. Warszawa: Wydawnictwo IFiS PAN.

Levy, A., Solovay, R. M. (1967). Measurable Cardinals and the Continuum Hypothesis. *Israel Journal of Mathematics*, *5*(4), 234–248.

Maddy, P. (1988a). Believing the Axioms. I. *Journal of Symbolic Logic*, *53*(2), 481–511.

Maddy, P. (1988b). Believing the Axioms. II. *Journal of Symbolic Logic*, *53*(3), 736–764.

Maddy, P. (1993) Does V Equal L? *Journal of Symbolic Logic*, *58*(1), 15–41.

Maddy, P. (2000). Does Mathematics Need New Axioms? *The Bulletin of Symbolic Logic*, *6*(4), 413–422.

Mancosu, P. (2008). Mathematical Explanation: Why It Matters. In: P. Mancosu (Ed.), *Philosophy of Mathematical Practice* (pp. 134–150). Oxford: Oxford University Press.

Mancosu, P. (2018). Explanation in Mathematics. *The Stanford Encyclopedia of Philosophy*. Retrieved from: https://plato.stanford.edu/archives/sum2018/entries/mathematics-explanation/

Marciszewski, W. (2018). Does Science Progress Towards Ever Higher Solvability Through Feedbacks Between Insights and Routines? *Studia Semiotyczne*, *32*(2), 153–185.

McCarty, D. C. (2004). David Hilbert and Paul du Bois-Reymond: Limits and Ideals. In: G. Link (Ed.), *One Hundred Years of Russell's Paradox* (pp. 517–532). Berlin, New York: Walter de Gruyter.

Mordell, L. (1959). *Reflections of a Mathematician*. Montreal: Canadian Mathematical Congress.

Murawski, R. (1984). G. Cantora filozofia teorii mnogości. *Studia Filozoficzne*, *11–12*(8–9), 75–88.

Paris, J., Harrington, L. (1977). A Mathematical Incompleteness in Peano Arithmetic. In: J. Barwise (Ed.), *Handbook of Mathematical Logic* (pp. 1133–1142). Amsterdam: North-Holland.

Pogonowski, J. (2006). Projekt logiki infinitarnej Ernsta Zermela. *Investigationes Linguisticae*, *XIV*, 18–49.

Purkert, W. (1989). Cantor's Views on the Foundations of Mathematics. In: D. E.

Rowe, J. McCleary (Eds.), The History of Modern Mathematics (vol. 1, pp. 49–65). San Diego: Academic Press.

Rav, Y. (1999). Why Do We Prove Theorems? *Philosophia Mathematica*, *7*(3), 5–41.

Rota, G.-C. (1997). The Phenomenology of Mathematical Proof. *Synthese*, *111*(2), 183–196.

Solovay, R. M. (1995). Introductory Note to *1970a, *1970b, *1970c. In: S. Feferman (red.), *Kurt Gödel: Collected Works: Volume III* (pp. 405–420). Oxford: Oxford University Press.

Steel, J. R. (2000). Mathematics Needs New Axioms. *The Bulletin of Symbolic Logic*, *6*(4), 422–433.

Tieszen R. (1998). Gödel's Path from the Incompleteness Theorems (1931) to Phenomenology (1961). *The Bulletin of Symbolic Logic*, *4*(2), 181–203.

Wang, H. (1987). *Reflections on Kurt Gödel*, Cambridge: MIT Press.

Wang, H. (1996). *A Logical Journey. From Gödel to Philosophy*. Cambridge: MIT Press.

Woodin, W. H. (1999). *The Axiom of Determinacy, Forcing Axioms and the Nonstationary Ideal*. Berlin, New York: de Gruyter.

Woodin, W. H. (2001). The Continuum Hypothesis. Parts I and II. *Notices of the AMS*, *48*(6–7), 567–576, 681–690.

Wójtowicz, K. (2002). *Platonizm matematyczny. Studium filozofii matematyki Kurta Gödla*. Tarnów: Biblos.

Originally published as "Kategoria wyjaśniania a filozofia matematyki Gödla". *Studia Semiotyczne*, *32*(2), 107–129, DOI: 10.26333/sts.xxxii2.07. Translated by Martin Hinton.

Studia Semiotyczne—English Supplement, t. XXX (2019), s. 107–127
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/stsen.xxx.06
Article

Paweł Stacewicz*

UNCOMPUTABLE NUMBERS AND THE LIMITS OF CODING IN COMPUTER SCIENCE[1]

Summary: The description of data and computer programs with the use of numbers is epistemologically valuable, because it allows to identify the limits of different types of computations. This applies in particular to discrete (digital) computations, which can be described by means of computable numbers in the Turing sense. The mathematical fact that there are real numbers of a different type, i.e. uncomputable numbers, determines the minimal limitations of digital techniques; on the other hand, however, it points to the possibility of the theoretical development and physical implementation of computationally stronger techniques, such as analogue-continuous computation. The analyses presented in this article lead to the conclusion that physical implementations of unconventional (non-digital) computations require the occurrence of actually infinite quantities in nature. Although some arguments of theoretical physics support the physical existence of such quantities, they are not definitive.

Keywords: numerical coding, computable numbers, uncomputable numbers, Turing machine, digital computation, analogue computation, infinity.

* Warsaw University of Technology, Faculty of Administration and Social Sciences. E-mail: p.stacewicz@ans.pw.edu.pl. ORCID: 0000-0003-2500-4086.

[1] I thank two anonymous reviewers for valuable comments and advice which significantly improved the original version of the text. I also thank the professors Witold Marciszewski and Andrzej Biłat for fruitful, editorial discussion of the text. For all errors, ambiguities and shortcomings remaining in the work I take responsibility and I apologize in advance for them to all readers.

From the point of view adopted in this work, computational objects, in particular computer programs, mediate between the mathematical sphere of numbers and physical reality. For example: a sound playing program operates on numerical representations of acoustic waves, and its instructions cause, due to the appropriate design of the computer, real physical vibrations of air molecules. What is more, this and any other program can be analysed on two levels, i.e. as an object of two types: on the one hand as a series of symbols that can be reduced to numbers, and on the other—as a strictly defined system of physical states of the machine (which, after running the program, cause regular changes of its subsequent states).[2] Due to the indicated correspondence, many computer-related issues can be resolved by referring to the properties of numbers—numbers that according to a particular, machine's specific, model of computation (e.g. digital or analogue) correspond to the data, texts and results of the programs.

In this work I shall focus on programs for digital machines. They are described theoretically by means of the Turing model of computation (universal Turing machine), and speaking "numerically", using computable numbers in the sense of Alan Turing. Referring to certain properties of computable and uncomputable numbers, in particular the fact that the digital representations of uncomputable numbers are actually infinite, I shall determine the theoretical reasons for the existence of computational limitations of such programs. I shall also discuss the possibilities of overcoming these limitations by means of computational techniques that (theoretically) allow the processing of signals described using uncomputable numbers in the Turing sense. The presented text is for the most part a review. However, it contains a number of the author's interpretations of the results of computer science and its mathematical foundations research (e.g. results of A. Turing and G. Chaitin), in particular interpretations regarding the infinite nature of uncomputable numbers and codes considered in theoretical computer science.

[2] Some philosophers of computer science speak directly—adopting an ontological rather than epistemological attitude—about the dual, i.e. abstract-physical, nature of computer programs (Moor, 1978; Colburn, 2000; see also Angius & Turner, 2013).

1. Numbers, Computing and Numerical Coding

The most important, and oldest, idea that resulted in the creation of computers and then computer science is the idea of numerical coding.[3] Behind it is the belief that the world of numbers (maybe even only natural ones) and relatively simple operations on them (such as comparing, adding or dividing) is rich enough to represent various aspects of the real world.

In modern computing, *numerical coding,* understood as describing data processed by computers using numbers,[4] is a common and perhaps theoretically necessary activity.[5] It is already present at the level of initial formalization of some tasks, when the objects appearing in these tasks (e.g. text, sound or graphic) are described by means of numbers, specially selected and included in appropriate structures. For example: characters processed by text editors are assigned specific numbers (according to e.g. ASCII code), while images displayed on monitors are often coded in the form of a sequence of numbers that determine the coordinates and colours of points on the raster matrix. At the lowest level of intra-computer structures, the relevant codes are created automatically, thanks to specially designed programs (e.g. compilers). Most importantly, however, in

[3] Its oldest manifestation was probably the philosophy of the ancient Pythagoreans, which postulated reducing all fragments of reality to some kind of numbers (summarized in the short slogan that "everything is a number"). In modern philosophical thinking, especially in the context of computer science philosophy, Pythagorean ideas are revived, which some call neopythagoreanism. This is due to a kind of feedback: Pythagorean ideas contributed to the emergence of computer science, and its successes, among others in the field of simulation of physical phenomena by means of operations on computer-represented numbers, strengthen the Pythagorean view of the world (Krajewski, 2014).

[4] In the computer science context, the term "describing data processed by computers using numbers" usually has a syntactic rather than an abstract sense. This means that it is about coding data using symbolic (and physical) representations of numbers, e.g., zero-one sequences. In the present text I shall also refer to the abstract (strictly mathematical) properties of numbers and their sets, such as the continuity of a set of real numbers. In the case of insufficient context, however, I shall signal whether in the given place it is about the abstract or syntactic dimension of the concept of number (writing e.g. that it is about decimal expansion of a number).

[5] See the online discussion on the Cafe Aleph blog that resulted from the creation of this work (Stacewicz, 2018b).

mathematical terms, aside from the physical design of the computer and the physical processes of signal processing, they can be represented numerically, for example in binary form.

The last five words of the previous paragraph indicate that I understand the term "numerical coding" widely in this work. In particular, I understand it more broadly than the term "digital coding", which I reserve for the way of representing information in digital computers, which are machines with discrete states operating on binary signals. I take the broad term "numerical coding" to be reasonable, because computer science, generally conceived, considers a wider class of machines than the digital. This broader class includes analogue circuits that allow (at least theoretically) operation on continuous signals described by real numbers,[6] as well as quantum computers for which the basic unit of information is the q-bit, mathematically defined using complex numbers.[7]

The concept of numerical coding is closely related to the key computer science concept of *computing*. In the context of problem solving, it means the mechanical implementation of the process of determining the value of the function, which assigns its specific solutions to the input of the problem (solutions for specific data).[8] If the data are numerically encoded, then the arguments and values of this function are those types of numbers (e.g. natural or real), which are allowed by the coding method appropriate for a given machine. This is determined by the appropriate model of computation (e.g. digital or analogue). Let us also say that the computer description, not purely mathematical, of the calculated function

[6] See works by Shannon (1941) and Rubel (1993).

[7] I also use the term "numerical coding" in another work (Marciszewski & Stacewicz, 2011, pp. 75–77). A similar conceptual convention is found in Krajewski, who does not use the term "numerical coding", but distinguishes digitization as one of the types of coding (although the most common), fundamentally different from data coding in analogue circuits processing signals described by real numbers (Krajewski, 2014).

[8] Historically, the first mature considerations for solving problems using calculations (computations), i.e. mechanical operations on physical equivalents of numbers, are due to G.W. Leibniz. For the modern concept of computing, the following ideas and achievements are particularly important: the design of a calculating machine (performing the four basic arithmetic operations), the invention of a binary arithmetic system, the design of a machine operating on binary encoded numbers, as well as the concept of a universal symbolic language (*lingua characteristica*) and coupled with it a reliable calculus (*calculus ratiocinator*). See the work by Trzęsicki (2006).

is either the program text (if the machine accepts programs written in a certain programming language) or the connection diagram between the elementary systems of the machine (if the machine is physically programmed like analogue systems or the first digital computers).

Since the vast majority of today's computers perform digital computations, in the following paragraphs I shall take a closer look at the "numerical" characteristics of the tasks entrusted to them. In particular, I shall consider the question as to whether the numerical codes desired in their description must be finite, or if sometimes it is necessary to refer to the concept of infinite code.[9]

At first glance, all the codes involved are finite, and thus reduced to natural numbers. This suggests the observation that the data entered into the digital computer have a finite representation, and the programs used to process them are finite sequences of instructions that, when encoded in binary form, can be interpreted as natural numbers. A deeper reflection on the functions of digital computers, however, leads to the statement that the theoretical analysis of the capabilities of these computers must refer to the concept of infinite code (even if such codes cannot be implemented inside real digital machines). Two possible contexts of reference should be distinguished.

First, in the case of many real problems (e.g. in the fields of dynamics or mechanics), the results obtained for specific input data can be expressed in *irrational numbers*, those with infinite and irregular expansions (e.g. decimal). This happens, for example, when a given problem is formu-

[9] The concept of infinite code—that is, the result of the coding process that has (actually) infinite length—is a non-standard concept that goes beyond the standard theory of computation, expressed e.g. in terms of Turing machines. However, in modern computer science methodology, which also includes some non-standard models of computation, this concept is used—e.g. to refer to the infinite length of program codes or the infinite tape of the Turing machine, which is completely filled with data (Ord, 2002, p. 17; Ord, 2016, p. 146; Mycka & Olszewski, 2015, pp. 58–59). Let us emphasize, however, that this concept makes sense when one makes an (even working) assumption of the possibility of going beyond the traditional Turing model of computation. The use of the concept of infinite code is justified in the present work, because later in it (especially in section 4) I shall analyse the possibility of the physical implementation of non-Turing computations, also those that include infinitistic elements. Regardless of this intention, in this section I show how (general) analysis of problems that we would like to solve traditionally (i.e. digitally) leads to the necessity of at least a critical consideration of non-traditional (i.e. infinite) codes.

lated mathematically using a certain equation (e.g. differential) and the root of this equation is an irrational number (such as $\sqrt{2}$, π or e). In this case, the result is *de facto* represented by an infinite number. Let us first note, before explaining in more detail in section 3, that the most troublesome situation occurs when we deal with this kind of irrational number, which is uncomputable in the sense of Turing. Secondly, however, and crucial for further analysis, each more complex programming task has an *infinitystic* structure. This means that the set of its initial data, and sometimes also the set of its potential results, is unlimited. As a simple example, let's consider the problem of determining the roots of quadratic equations $ax^2 + bx + c = 0$, where the range of possible a, b, c coefficients to enter is unlimited. In the case of this problem, there is, despite an unlimited field, a finite method of finding the x values sought, which is the commonly known "delta" algorithm. There is also a finite program (many), which for any input data (i.e. a system of coefficients a, b, c) allows, in a finite number of steps, the generation of the correct result. This program must be treated as a general (computer) solution to the problem posed, a solution which corresponds to the finite numerical code of the program (in short: a certain number).[10]

Unfortunately, for other problems with an unlimited input data domain, the numerical code of the general solution—which is a digital record of all possible pairs <INPUT, RESULT>, or in other words, the function that assigns the results—must remain infinite. This happens when there is no finite program to solve the problem. If such a program exists, it is a form of encoding the set of the given pairs in the shape of a procedure that generates correct results (for all possible input data) which is "intelligible" for a digital machine. The code of such a procedure corresponds to a natural number (written e.g. as a sequence of zeros and ones). If such a program does not exist, it must be assumed that the overall solution to the problem corresponds to some uncomputable number in the Turing sense (i.e. a certain special irrational number with infi-

[10] Let us emphasize here that, although the question of the infinite domain of input data may be irrelevant from the point of view of solving the algorithmic task, the fact that this solution applies to an unlimited number of input data determines its strength. It is in a way a universal solution (similar to mathematical theorems, it applies to an infinite number of special cases). In some situations, however, the infinite field can lead to trouble—more on this in the main text (see also Stacewicz, 2015).

nite expansion, which no digital machine can calculate; see further in section 2).

In the context of solving problems by computing that interests us here, infinite numerical codes can, therefore, occur at two levels: 1) at the code level of the exact individual result, 2) at the code level of the overall problem solution. In both cases, it may happen that the appropriate code is in the form of an uncomputable number, and then—as we shall see in section 3—the method sought to solve the problem lies beyond the limits of the possibilities of digital coding (which does not, however, exclude the existence of such a method that would be implemented on machines of types other than digital).

2. Uncomputable Numbers in the Turing Sense

The *uncomputable numbers* highlighted in the title of this article were defined by Alan Turing in his work from 1936 entitled *On Computable Numbers, with an Application to the Entscheidungsproblem.* He defined them as such irrational numbers, whose decimal representation cannot be determined with any given accuracy, by any system for mechanical calculations, today called the Turing machine.[11] In the modern style, we would say that these numbers are indeterminate by means of algorithms for digital machines, and therefore those for which there are no finite computer programs that allow step by step calculation of the subsequent digits of their decimal or other representations (although such representations are strictly defined, see Stacewicz, 2012). For example: the irrational number e does not have the above properties, because it is relatively easy to generate successive digits of its expansion by means of a program calculating successive subtotals of the appropriate series (remember that $e=\sum_{n=0}^{\infty}\frac{1}{n!}=1+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+\ldots$). Therefore, it is not an uncomputable number, although it is characterized by irrationality.

Unlike the irrational number e, uncomputable quantities in the Turing sense are defined in a way that excludes the possibility of their successive approximation using Turing machines or equivalent computational mech-

[11] It is worth adding that Turing first gave the exact definition of a set of computable numbers (numbers whose decimal notation can be determined definitively or with any given accuracy using a finite program for a Turing machine), and then proved the existence of real numbers of another type (see further in main text), or uncomputable numbers (Turing, 1936).

anisms. This is determined by Turing's original reasoning, which, after defining computable numbers, proved that there are real numbers of another type, and then specified a set of non-computable numbers as the complement to the set of computable numbers in the set of real numbers. I shall present this reasoning in a sketchy and demonstrative way—limiting it to real numbers in the range (0,1).[12] The starting point of the argument is that the result of the operation of each Turing machine for specific input data—a machine generating series of digits from the set {0,1, ..., 9}—is clearly represented by a certain real number from the interval (0,1). It is such a number whose decimal expansion is the same as the finite or infinite sequence of digits generated by the machine.

Due to the fact that each machine, together with the input data, unambiguously defines a unique string of symbols (representing its program and the initial content of the tape), each of them can be assigned a unique number, and all machines can be set into an infinitely countable sequence. According to the order in this sequence, you can then set all digit sequences generated by subsequent machines. These sequences form an infinite countable set and unambiguously designate concrete computable numbers in the range (0,1). These are numbers with decimal expansions identical to the subsequent sequences.

Having the above-mentioned sequence list, one can ask if there is such a sequence S on it, that its n-th digit differs (e.g. by 1) from the n-th digit of the n-th sequence on the list (if the n-th sequence is long enough). The postulated S sequence cannot appear in the list because it differs (by at least one number) from each of the sequences in the string. Therefore, it differs from any sequence generated by any machine. Therefore, this sequence must specify a number from the range (0,1) that no machine can generate, i.e. a real uncomputable number (Turing, 1936; Marciszewski & Stacewicz, 2011).[13]

---

[12] In the presented reasoning, Turing skilfully used the diagonal technique, which was used for the first time by G. Cantor in proof of the uncountability of the set of real numbers.

[13] Let us also note that the procedure of determining the S sequence proposed above is inefficient (although theoretically allowed), because due to the insolvability of the Turing machine halting problem (in the quoted Turing work we will find the appropriate proof), we do not know which of the machines generating the sequences on the list stops, and which does not (moreover: in the second case we do not know whether the machine head will not "turn" in any cycle and will not

The first exact definition of uncomputable numbers of some kind was given by the modern mathematician, Gregory Chaitin. These are Omega numbers, which for a universal Turing machine of a given type (i.e. machines with a certain number of states and symbols of the alphabet) determine the probability that a randomly selected program of operation of such a machine will stop.[14] Let us also clarify that by the program of operation is meant here the initial content of the universal machine tape, which consists of a properly coded program of the simulated machine and its initial data.[15] It refers, therefore, to the input data of the universal machine, which however strictly define its subsequent activities (the universal machine implements the program of the concrete machine for the specific data). Since the construction given by Chaitin is quite complex and serves to determine the formula for the mentioned probability (Chaitin, 1993; Chaitin, 2005), I propose here a conceptually simpler definition of another uncomputable number. I shall keep Chaitin's original idea, which refers to the issue of the halting of Turing machines.[16]

The starting point of the definition is to prepare an ordered list of programs for the universal Turing machine of certain type. As in the case of Chaitin's construction, by program I understand the initial content of the universal machine tape (including the program code of a specific machine and its input data). Since the aforementioned list is countably infinite,[17] the programs on it (with data) can be numbered as p_1, p_2, p_3 etc.

---

change the above-mentioned nth digit). We shall refer to the issue of halting further by defining an uncomputable number L.

[14] The subject literature often mentions one Omega number (see, e.g., Trzęsicki, 2006a, pp. 125–126). However, this is confusing, because for each universal Turing machine (there are infinitely many such machines) there is a separate Omega number, having a different symbolic representation.

[15] In addition to the program thus understood, each universal machine has its unique (defining it) "executive" program, which determines the way of implementation of each program placed on the tape (it regulates, among other things, how the machine head moves between the simulated machine program code and its input).

[16] Remember that this issue is expressed by the question about the existence of such a (diagnostic) Turing machine, which for each other Turing machine and each of its input data would be able to unequivocally decide whether this particular machine will stop working for this particular input or whether it will work forever.

[17] It is infinite, because due to the infinite length of the universal machine tape, there are infinitely many input data that can be put on it (despite the finite number of alphabet symbols and the finite number of states of the simulated specific machines).

Referring to this list, one can define the following binary number L in the range (0,1) : $L = 0$, $b_1b_2b_3$ ..., where bit $b_i = 1$ if the p_i program stops, and $b_i = 0$ if the p_i program does not stop (where $i \in N$).[18]

Note that the number L is strictly defined—because programs that define its subsequent bits either stop or fail. However, it is not computable, i.e. algorithmically determinable—because determining the values of subsequent bits the issue of halting cannot be algorithmically resolved in finite time. This example again shows that Turing's uncomputability is strongly associated with infinity. The number L has an infinite expansion, which is indeterminate by the finite program (indeterminate in the sense that the calculation of some of its digits would take infinitely long).

Developing the "infinity thread" in a more general context, it must be stated that all numbers that are uncomputable in the Turing sense are characterised by actual infinity (not potential[19]). Each of their symbolic representations (e.g. decimal) contains an infinite number of digits, which must be understood as an infinite whole, impossible to gradually generate, digit by digit, using any finite program (for a digital machine).[20]

---

[18] As Chaitin notes, the issue of choosing the right list, i.e. how to order the set of programs, is extremely important. It should be emphasized that it is important not only in the case of defining Omega numbers (in their case Chaitin gave a special way to specify the list), but also in the definition of another type of uncomputable numbers (Chaitin, 1993). One of the anonymous reviewers of this paper rightly stated that the type of the number L specified in the main text (computable or non-computable) depends on how the p_i program set is ordered (i.e. how the list is compiled). In particular: computable numbers (such as 2/3) can be obtained for certain orders. To solve this problem, the above-mentioned definition of Chaitin's list can be adopted. Notwithstanding the above explanations, it should be emphasized, however, that the number L is defined in such a way that even if the list in its definition causes its computability, this definition alone does not allow one to state that computability on the basis of any operations implemented by Turing machines. This is because the basis of the definition is the halting problem, and its undecidability makes it impossible to determine (in advance) which programs on the list stop and which do not. In short: perhaps for a certain list of programs the number L is computable, but we, using only the Turing machine operations, are unable to determine it.

[19] For the distinction between potential and actual infinity, see Murawski (2014). Also worth noting is the text by Witold Marciszewski on infinity (2012).

[20] The actual infinity of an uncomputable number is well illustrated by the following metaphor: if some super-algorithmic Divine Mind wanted to share with us the knowledge of an uncomputable number X, it would have to reveal it to us

Let us finish by explaining that the class of uncomputable numbers in the sense of Turing is extremely extensive, because it has the cardinality of the continuum, and therefore is equinumerous with the set of real numbers. In contrast to it, the class of computable numbers, i.e. those that are algorithmically determinable using Turing machines, has the cardinality aleph-null, i.e. is equinumerous with the set of natural numbers.[21] This disproportion between the infinities of sets of computable and uncomputable numbers seems surprising: everything that Turing machines can generate turns out to be "a drop in the ocean of uncomputability."

3. Minimum Limitations of Real Digital Codes

By real digital codes I understand here the numerical codes of the actual programs that can be physically implemented, which programs in a finite way represent functions that associate input data and results of computations. Due to the computational equivalence of (idealized) digital computers and Turing machines,[22] the results of these computations are always digital representations of some computable numbers in the Turing sense (or fragments of them, if the number has an infinite expansion).

Due to this equivalence, the general limitations of real digital codes—limitations that must be met by all programs for all digital machines—can be determined within the Turing model of computation, which is in

---

in its entirety, an infinite whole, but would not be able to provide a concise algorithmic rule describing it in a finite way. This is a casual paraphrase of Chaitin's remarks (Chaitin, 1998, pp. 54–55). I write more about the difference between the types of infinity of the computable numbers (potential infinity) and the uncomputable numbers (actual infinity) in another work (Stacewicz, 2018a, pp. 180–181).

[21] This is due to the fact that all machines that generate unique strings of symbols that make up symbolic representations of computable numbers can be numbered and set into an infinite string. The set of uncomputable numbers must have the cardinality of the continuum, because it is defined as the difference of the set $\mathbb{R}$ (with the cardinality of the continuum) and the set of computable numbers (with the cardinality *aleph-null*).

[22] More precisely, each program of a certain digital machine (regardless of the technical details of its design) can be translated into the Turing machine program, in particular the universal machine program. Despite this, due to the purely physical limitations of real digital machines (not ideal, but real), not all tasks "feasible" for a UTM are feasible for them. This topic will be developed further in the main text of this section.

the form of an abstract universal machine, called the universal Turing machine (UTM).[23] More precisely, if the solution to a certain problem cannot be coded in the form of a program for a UTM, it also cannot be considered an executable procedure for a certain digital machine.[24] Which does not mean—we must add—that it cannot be specified in the form of a procedure for a machine of another type, e.g. analogue.

From the point of view of these considerations, the key role here is played by the issue of determining uncomputable numbers, and more precisely their subsequent digits, which constitute their symbolic representations. Such numbers have correct definitions, their subsequent digits (e.g. 0 and 1) are precisely defined, and yet there is no program for the Turing machine that would allow such numbers to be determined in any finite length of time. Thus, the functions corresponding to individual uncomputable numbers—functions that bind the given accuracy (e.g., the number of the last desired digit of the decimal number expansion) to the corresponding fragment of the number—determine the limits of the digital coding. If the general solution to a given problem is reduced to this kind of function, then this solution cannot be digitally coded. To put it another way: if, for a certain problem P, each numerical code of a function that binds its input data and results corresponds to a certain uncomputable number, then this problem lies (then and only then) beyond the limits of the possibilities of digital coding. In this way, i.e. by explaining "numerically" the issue of computational unsolvability of some problems, we gain some new insight into both the reasons for, and the hypothetical possibilities of, overcoming Turing's uncomputability.

Limitations set by uncomputable numbers, and more precisely by the functions associated with them, generating their symbolic representations, should be treated as minimum limits, independent of the physical characteristics of digital machines. This statement results from the fact that the UTM machine is computationally equivalent not to physical digital machines, but to theoretical computers, with infinite memory resources and

[23] Let us remind that a universal Turing machine is a machine that, thanks to a specially selected program defining it, is able to simulate the operation of any particular Turing machine (Harel, 2000, p. 252).

[24] A wide range of uncomputable problems in the Turing model are described, for example, by Harel (2000, pp. 201–224). Gödel also mentions some important meta-mathematical problems of this type (1995/2018, p. 13).

an arbitrarily long, although finite, operating time.[25] It means that a UTM machine is able to "perform" more tasks than physical digital machines of a certain type (e.g. machines with a maximum RAM of 8 MB). Hence the conclusion that the limitations of real physical computers and the digital codes controlling them are in fact greater than the limitations of idealized machines, i.e. Turing machines. The limitations of the latter are therefore the "mathematical minimum", covering all digital computers.

Let's return to the properties of uncomputable numbers. Remember from section 2 that all representations of such numbers are characterized by infinity. These representations are in fact infinite wholes—that is, infinite sequences of symbols, which are not determined by any finite rule, having the form of a finite program for a Turing machine. From this perspective, the actual infinity of the numbers that would have to code the solutions of some problems should be considered the mathematical "cause" of Turing's uncomputability of these problems.

Due to the previously indicated correspondences between specific numbers of this type and digitally uncomputable problems (e.g. the previously determined number L corresponds to the halting problem), and the fact that the set of uncomputable numbers has the cardinality of the c o n t i n u u m , the conclusion is that uncomputable problems in the Turing sense are infinitely many, and moreover, that there are many more than there are computable ones (whose set, like the set of computable numbers, has the cardinality aleph-null). This is a conclusion, not a supposition, because each uncomputable number has at least one unsolvable problem, consisting in determining any fragment of its digital representation.

It can, of course, be argued that the infinite continuum of digitally uncomputable problems contains a relatively small number of practically relevant issues. For example, even the halting problem—as it concerns all Turing machines, and not just some of their highlighted subsets—can be considered too wide and thus insignificant from a practical point of view. However, the extremely practical point of view seems illusory. It is difficult to be sure that solutions to problems that do not translate directly into applications do not conceal practically significant consequences

[25] In the UTM model, an infinite tape is responsible for the potentially infinite memory resources and potentially infinite operating time (Stacewicz, 2018a).

(which at a given stage of the development of science and technology we do not yet know).[26]

Before moving on to the next section, devoted to alternative techniques to Turing computations, it is worth paying attention to one more feature of uncomputable numbers. In relation to the set of numbers available for Turing machines, i.e. computable ones, they are elements that, going beyond this set, allow it to be "expanded" to the form of a set of real numbers. This in turn suggests that there may be such computational techniques that refer to the theory of real numbers (and further: to some results of mathematical analysis), and, in the implementation layer, allow for operation on the physical equivalents of some or all real numbers. We'll look at the possibilities of such techniques in the next section.

4. Can There Be Effective Implementation of Non-Digital Codes?

Due to the properties of digital computers,[27] all codes representing data, programs and the results of these devices are subject to certain minimum restrictions, determined within the Turing model of computation. In fact, these restrictions consist in the inability to "go beyond" a set of computable numbers in the sense of Turing.

In connection with the above, the question arises as to whether there are any computing machines, other than digital, that would be able to operate on real, non-computable codes, i.e. certain physical representations of non-computable numbers in the Turing sense. If such machines actually existed, they could, firstly, solve problems whose only available general solutions are encoded with uncomputable numbers, and secondly, they could generate results that are such numbers (or represented by them). The computing power of such machines would, therefore, be greater than the power of digital devices.

---

[26] To justify the belief in the practical significance of any uncomputable problems, one can rely on somewhat breakneck but suggestive reasoning by analogy. Well, just as in the set of real numbers, you cannot omit (without prejudice to their mathematical utility) uncomputable numbers (because their existence gives the set R the property of continuity), so in the set of all problems you cannot miss out the set of uncomputable problems. This reasoning would require further development, which is why we only signal it in the footnote.

[27] Remember that this is about computational equivalence of (idealized) digital computers and Turing machines.

From the point of view of pure theory, such machines exist, and the general principles of their operation are determined by various models of hypercomputation—so-called because of their proper potential for expanding the capabilities of the UTM machine (Copeland, 2002). These include, among others: infinity models—allowing for an infinite number of operations (computations) in a finite time (Shagrir, 2004); non-deterministic models—describing computations initiated and/or randomly controlled (Deutsch, 1985); and analogue—allowing processing of continuous signals, mathematically described using real numbers from a specific range (Mycka & Piekarz, 2004). It is worth emphasizing that the idea of non-digital coding manifests itself most fully in the case of computations of the last type, i.e. analogue, because their theory gives the opportunity to operate on quantities (codes) from the entire continuum (and not on codes described by specific uncomputable numbers).[28]

Theoretical proposals of computations of one or another type obviously do not prejudge the issue of their physical feasibility. This issue is negatively resolved by the Church-Turing hypothesis, which in one version states that "a function is effectively computable if and only if it is computable using the universal Turing machine" (Harel, 2000, p. 240).[29] In the context of coding, this wording can be interpreted so that the only effectively processable codes are data acceptable to, and possible to generate by, the UTM machine, i.e. digital (discrete) codes. From this perspective, therefore, all codes, regardless of their theoretical description, are practically reducible to digital codes—which depends on, among other things, the fact that there is always the possibility of approximating them using digital equivalents. Considering the fact that the UTM model is theoretical and defines more computational constraints than their real possibilities, the conclusion of the hypothesis can be described in a different way. The UTM model sets absolutely minimal coding limitations in computer science.[30] In other words: all real computations—regardless of

[28] It is also worth adding that analogue techniques remain the closest to the practice of computer science—both for historical reasons (because analogue machines were already being constructed in the 1930s) and from the perspective of modern research (Mycka & Piekarz, 2004; Shannon, 1941).

[29] I treat the quoted wording as a hypothesis, because I do not prejudge whether only Turing computations (implemented in practice by digital machines) are effectively physically feasible.

[30] In the previous section, in the fourth paragraph, I also explained that these are the minimum theoretical limitations of digital techniques.

the theoretical model that describes them—must be subject to the restrictions set out in this very close to model practice (i.e. UTM). The limitations of alternative designs, e.g., analogue models, are simply broader.

The most serious arguments for the truth of the Church-Turing thesis, and therefore also for the existence of the above restrictions, refer to the concept of infinity. The basic issue is the fact that uncomputable numbers—corresponding to solutions to certain problems—are characterized by actual infinity. Remember that it concerns their endless, irregular expansions, impossible to gradually generate, which as an infinite whole represent (digitally) a given number.

The determination of such representations, and thus the resolution of the corresponding problems, must require the use of physical, uncomputable quantities existing in nature. Embedding such natural carriers of uncomputability in a machine is necessary because it is known that the overall representations of uncomputable numbers cannot be coded or determined in a traditional way, i.e. using minimally "nature engaging" binary codes and operations.[31] In particular, all effective implementations of the abovementioned analogue techniques require the use of uncomputable physical quantities. This is due to the fact that both the specificity and strength of these techniques (i.e. their greater computing power than digital techniques) rely on the possibility of processing and generating quantities from a certain *continuum* (Mycka & Piekarz, 2004). This, however, would not be continuous were it not for the uncomputable quantities filling it.[32]

Therefore, the real problem of the existence of carriers of uncomputability in nature arises. Remember that their most problematic feature is their having physical, but in accordance with the theoretical properties of

[31] Binary codes and operations must also be physically implemented using one or other natural quantities (e.g. electrical pulses); the thing is, however, that in their case it is enough to use any physical quantities that are easily distinguishable (or even one recognizable quantity and the lack thereof). Thus, the degree of "engagement" of nature is minimal in their case.

[32] The same fact can be expressed by referring to the properties of real numbers, which are the mathematical equivalent of processed continuous analogue signals. Well, without uncomputable numbers, each range of real numbers (equivalent to the physical domain of analogue signals) has the cardinality *aleph-null*, so it is equivalent to a discrete set of natural numbers.

uncomputable numbers, actual infinity.[33] If such carriers existed, the set of practical computational codes would go beyond the set of digital codes. It would include codes that have direct roots in nature. Some of their components, at least, would simply be "calls" to natural phenomena that would return some uncomputable quantities directly and in whole. In particular, the theory of analogue-continuous computation initiated by Claude Shannon (Shannon, 1941) states that a complex analogue code may include elementary integration operations, whose continuous results (implemented in real time) must be obtained by measuring phenomena occurring in special physical systems (e.g. electronic integrators). And as I wrote above, for the continuity of the result set it is necessary for it to contain uncomputable quantities.

The existence of *uncomputable* natural phenomena—that is, those that cannot be described in terms of computable numbers and functions implemented by Turing machines—postulates certain physical theories. One particularly cited example is from Pour-El and Richards (1989). According to it, the three-dimensional wave described by a certain differential equation can obtain that can be expressed only by means of uncomputable numbers. John Doyle's proposals which indicate the inability to describe the processes of achieving equilibrium occurring in nature (e.g. thermodynamic) using computable functions fall into the same category (Copeland, 2002, p. 470). These and other examples seem to indicate the real existence of phenomena that we could treat as natural carriers of uncomputability. Let us remember, however, that empirical tests are responsible for the compatibility of physical theories with reality, which no finite number (again, an infinity problem!) can ever confirm with 100% certainty.

Suppose, however, regardless of the above objection of an epistemological nature, that physical carriers of uncomputable codes exist and can be used as part of one or other natural computations.[34] Despite this assump-

[33] The philosophical argument for the existence of infinite quantities in nature is contained in *Amor Infiniti. What philosophical intuitions lead to it?* (Marciszewski, 2012).

[34] I am thinking of computation designed by man, but involving significantly substrates and/or natural processes (e.g. quantum calculations or those performed using DNA molecules). The class of natural computation in a broader sense also includes: 1) computation inspired by observation of nature (e.g. implemented by artificial neural networks) and 2) processes occurring in nature, described in com-

tion, another problem arises concerning the possibility of *reading*, and thus knowing the obtained result. The problem is that in order to know the result, infinite accuracy in reading the entire uncomputable quantity is necessary.[35] It is necessary, because finite accuracy, which, after all, characterizes all real measuring instruments, would bring the uncomputable number desired in a given situation to the level of a finite computable quantity. Therefore, we would lose the expected effect of overcoming the limitations of digital computing. It can be argued that for some problems, it is enough for uncomputable quantities to be simply processed and not read—because the solution to the problem is some specific finite value that can be read (Stannett, 2003, pp. 121–123). The approach considered here is, however, about knowing the general solution to the problem (a function that associates all possible input data with the corresponding results), and this type of solution is encoded by an entire number that is uncomputable with actually infinite expansion. Therefore, the epistemic problem remains: without infinite accuracy of reading we cannot know such a solution.

To conclude: the actual infinity of uncomputable numbers means that the limitations of computational techniques suggested by the Church-Turing thesis—techniques that require the physical implementation of certain computational codes—can be overcome under at least two conditions: 1) the occurrence of infinite quantities in nature that can be recorded and processed, 2) the existence of a mental disposition for insight into actually infinite objects and their relations and methods (e.g. methods of defining). The second condition must be considered fulfilled—as evidenced by the actual infinity theories created by people, including theoretical models of computing on actually infinite quantities. The possibility of meeting the first condition seems, at least, problematic.

putational categories (e.g. intracerebral processes; see Kari & Rozenberg, 2008; Rozenberg, Back, & Kok, 2012).

[35] Such accuracy is necessary in the case of analogue techniques, which by definition operate on continuous quantities (two quantities in the continuous domain may differ from each other by any small amount).

REFERENCES

Angius, N., Turner, R. (2017). Philosophy of Computer Science. *Stanford Encyclopedia of Philosophy*. Retrieved from: https://plato.stanford.edu/entries/computer-science/

Chaitin, G. J. (1993). Randomness in Arithmetic and the Decline and Fall of Reductionism in Pure Mathematics. *Bulletin of the European Association for Theoretical Computer Science*, *50*, 314–328.

Chaitin, G. J. (1998). *The Limits of Mathematics*. Singapore: Springer.

Chaitin, G. J. (2005). Omega and Why Maths Has No TOEs. Retrieved from: https://plus.maths.org/content/os/issue37/features/omega

Colburn, T. R. (2000). *Philosophy and Computer Science*. Armonk, NY: M.E. Sharpe.

Copeland, J. (2002). Hypercomputation. *Mind and Machines*, *12*(4), 461–502.

Deutsch, D. (1985). Quantum Theory, the Church-Turing Principle and the Universal Quantum Computer. *Proceedings of The Royal Society of London A*, *400*, 97–117.

Etesi, G., Nemeti, I. (2002). Non-Turing Computations via Malament-Hogarth Space-Times. *International Journal of Theoretic Physics*, *41*(2), 341–370.

Gödel, K. (1995/2018). O pewnych zasadniczych twierdzeniach dotyczących podstaw matematyki i wnioskach z nich płynących. *Studia Semiotyczne*, *32*(2), 9–32.

Harel, D. (2000). Rzecz o istocie informatyki. *Algorytmika*. Warsaw: Wydawnictwa Naukowo-Techniczne.

Kari, L., Rozenberg, G. (2008). The Many Facets of Natural Computing, *Communications of the ACM*, *51*(10), 72–83.

Krajewski, S. (2014). Neopitagoreizm współczesny: uwagi o żywotności pitagoreizmu. In: M. Heller, S. Krajewski (Eds.), *Czy fizyka i matematyka to nauki humanistyczne?* (pp. 348–366). Kraków: Copernicus Center Press.

Leibniz, G. W. (1890). *Philosophische Schriften* (Vol. VII). Berlin: Weidmann.

Marciszewski, W. (2012). Amor Infiniti. Jakie doń prowadzą intuicje filozoficzne? Retrieved from: http://marciszewski.eu/?p=2955

Marciszewski, W., Stacewicz, P. (2011). *Umysł-Komputer-Świat. O zagadce umysłu z informatycznego punktu widzenia.* Warsaw: Akademicka Oficyna Wydawnicza EXIT.

Moor, J. H. (1978). Three Myths of Computer Science, *The British Journal for the Philosophy of Science*, *29*(3), 213–222.

Murawski, R. (2014). Nieskończoność w matematyce. Zmagania z potrzebnym, acz kłopotliwym pojęciem. *Zagadnienia Filozoficzne w Nauce*, *55*(2), 5–42.

Mycka, J. M., Piekarz, M. (2004). Przegląd zagadnień obliczalności analogowej. In: S. Grzegórski, M. Miłosz, P. Muryjas (Eds.), *Algorytmy, metody i programy naukowe* (pp. 125–132). Lublin: Polskie Towarzystwo Informatyczne.

Mycka, J. M., Olszewski A. (2015). Czy teza Churcha ma jeszcze jakieś znaczenie dla informatyki? In: P. Stacewicz (Ed.), *Informatyka a filozofia. Od informatyki i jej zastosowań do światopoglądu informatycznego* (pp. 53–74). Warsaw: Oficyna Wydawnicza Politechniki Warszawskiej.

Ord, T. (2002). Hypercomputation: Computing More Than the Turing Machine. Retrieved from: https://arxiv.org/ftp/math/papers/0209/0209332.pdf

Ord, T. (2006). The many forms of hypercomputation. *Applied Mathematics and Computation*, *178*(1), 8–24.

Pour-El, M. B., Richards, J. I. (1989). *Computability in Analysis and Physics*. Berlin: Springer.

Rozenberg, G., Back, T., Kok, J. N. (2012). *Handbook of Natural Computing*. Berlin-Heidelberg: Springer.

Rubel, L. (1993). The Extended Analog Computer. *Advances in Applied Mathematics*, *14*(1), 39–50.

Shagrir, O. (2004). Super-Tasks, Accelerating Turing Machines and Uncomputability. *Theoretical Computer Science*, *317*(1–3), 105–114.

Shannon, C. (1941). Mathematical Theory of the Differential Analyzer. *Journal of Mathematics and Physics*. *20*(1–4), 337–354.

Stacewicz, P. (2012). Co łączy umysł z teorią liczb? *Filozofia Nauki*, *79*(3), 111–126.

Stacewicz. P. (2015). Informatyczne kłopoty z nieskończonością. In: R. Murawski (Ed.), *Filozofia matematyki i informatyki* (pp. 310–327). Kraków: Copernicus Center Press.

Stacewicz, P. (2018a). Czy informatykom musi wystarczyć nieskończoność potencjalna? In: R. Murawski, J. Woleński (Eds.), *Problemy filozofii matematyki i informatyki* (pp. 177–190). Poznań: Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Stacewicz, P. (2018b). O teoretycznej (nie)zbędności kategorii liczby w informatyce i jej metodologii. Retrieved from: http://marciszewski.eu/?p=999

Stannett, M. (2003). Computation and Hypercomputation. *Minds and Machines*, *13*(1), 115–153.

Trzęsicki, K. (2006). From the Idea of Decidability to the Number Omega. *Studies in Logic, Grammar and Rhetoric*, *22*(1), 73–142.

Trzęsicki, K. (2006). Leibnizjańskie inspiracje informatyki. *Filozofia Nauki*, *55*(3), 21–48.

Turing, A. M. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *2– 42*(1), 230–265.

Originally published as "Liczby nieobliczalne a granice kodowania w informatyce". *Studia Semiotyczne*, *32*(2), 131–152, DOI: 10.26333/sts.xxxii2.08. Translated by Martin Hinton.

Studia Semiotyczne—English Supplement, t. XXX (2019), s. 129–150
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/stsen.xxx.07
Article

Marek Lechniak,* Andrzej Stefańczyk **

ARGUMENTATION STRATEGIES IN ARISTOTLE' S THEORY OF RHETORIC: THE APPARENT ENTHYMEME AND THE REFUTATIVE ENTHYMEME[1]

Summary: In the *Organon*, Aristotle distinguished two types of reasoning: analytical and dialectical. His studies on analytical reasoning in the *Prior and Posterior Analytics*, earned him the title of the father of formal logic. According to Chaim Perelman, modern logicians have failed to see the fact that Aristotle's considerations on dialectical reasoning in the *Topics*, the *Rhetoric* and the *Sophistical Refutations* made him also the father of the theory of argumentation. This article attempts to answer this diagnosis. Our aim is to prove Perelman's thesis on the homogeneity of Aristotle's concept of theoretical and practical syllogism. The key concept in this proof is that of the enthymeme. In the article, we will try to answer the question of what place the enthymeme occupies in Aristotle's theory of rhetoric and confront it with the concept of a syllogism. We will also outline the structure of argumentation that makes use of the enthymeme, and present

* The John Paul II Catholic University of Lublin, Faculty of Philosophy. E-mail: marek.lechniak@kul.pl. ORCID: 0000-0002-0768-7963.

** The John Paul II Catholic University of Lublin, Faculty of Philosophy. E-mail: astefanczyk@kul.pl. ORCID: 0000-0001-5621-0777.

[1] The project is funded by the Minister of Science and Higher Education within the program under the name "Regional Initiative of Excellence" in 2019-2022; project number: 028/RID/2018/19.

two types of enthymemes discussed by Aristotle: the apparent enthymeme and the refutative enthymeme.

KEYWORDS: argumentation, enthymeme, syllogism, Aristotle's rhetoric, apparent enthymeme, refutative enthymeme, non-monotonic logics.

1. THE ENTHYMEME AS A SYLLOGISM

Perelman points out that just as Peter Ramus drew a line between modern rhetoric and the art of argumentation (defining rhetoric as "the art of speaking well, the eloquent and decorative use of language"), also contemporary formal logic disregards the argumentative role of rhetoric and completely neglects dialectical reasoning. Perelman considers these two approaches to be erroneous, both substantively (because they ignore the function of logic as a tool for studying reasoning in all forms) and historically, as Aristotle applied one theory to both analytical and dialectical reasoning (Perelman, 2002, p. 13).

In fact, Aristotle in his *Rhetoric* points out two logical ways of reasoning that organize the subject of discourse: the enthymeme[2](ἐνθύμημα) and the example (παράδειγμα). They are counterparts of a syllogism (deduction) and an induction as the methods by which we learn about the real world in philosophy and in science (*Rhet.,* 1356B 1–5), for "every belief comes either through deduction or from induction."[3] Due to the common modes of persuasion[4]—as Aristotle writes about enthymemes and examples—the speech and the speaker himself can be classified as "using either

[2] Unless marked otherwise, all citations from the *Rhetoric* come from *The Complete Works of Aristotle—Revised Oxford Translation*, Vol. 2, ed. by Jonathan Barnes, Princeton University Press 1984.

[3] In the original: ἅπαντα γὰρ πιστεύομεν ἢ διὰ συλλογισμοῦ ἢ ἐξ ἐπαγωγῆς (*APr,* 68b 13–14). Unless marked otherwise, all citations from the *Prior Analytics*, *Posterior Analytics*, *Topics* and *Sophistical Refutations* come from *The Complete Works of Aristotle—Revised Oxford Translation*, Vol.1, ed. by Jonathan Barnes, Princeton University Press 1984. The article uses the commonly accepted Bekker numbering.

[4] In the original: αἱ γὰρ πίστεις ἔτεχνόν ἐστι μόνον, τὰ δ' ἄλλα προσθῆκαι... (Rhet., 1354a 13–14).

enthymemes or examples."[5] The rationale for using one method or the other is that:

> induction is more convincing and clear: it is more readily learnt by the use of the senses, and is applicable generally to the mass of men, but deduction is more forcible and more effective against contradictious people (τοὺς ἀντιλογικοὺς ἐνεργέστερον).[6]

The way Aristotle writes about the enthymeme in the *Rhetoric* and the amount of space he devotes to it clearly show how important this concept was for him.

What is an enthymeme? Although Aristotle states that enthymemes are "the substance of rhetorical persuasion" (*Rhet.,* 1354a 14–15), he fails to give a precise definition of an enthymeme.[7] This failure, however, is only apparent. The definition of an enthymeme is not given explicitly, but it can be inferred from Aristotle's logical works (the *Prior and Posterior Analytics*, the *Topics*) and from the *Rhetoric*. It is in the *Rhetoric* in particular that the relation between an enthymeme and a syllogism is often emphasized,[8] which, combined with Aristotle's logical texts, makes it possible to identify what an enthymeme is.

In the *Prior Analytics* and the *Topics* (*Top.*, 100a, 25ff, 165a 1 ff.), we can find a definition of syllogism (deduction), which goes as follows:

> A deduction (συλλογισμός) is discourse in which, certain things being stated, something other than what is stated follows of necessity from their being so. I mean by the last phrase that it follows because of them and by this, that no further term is required from without in order to make the consequence necessary (*APr,* 24b 18–26).

This definition is so broad that it includes all forms of inference. On the other hand, when contrasted with another passage which says that "deduction is the more general; a demonstration is a sort of deduction (ἡ

---

[5] In the original: καὶ ῥήτορες ὁμοίως οἱ μὲν παραδειγματώδεις οἱ δὲ ἐνθυμηματικοί. (*Rhet.,* 1356b 27–28).

[6] *Top.* 105a 16–19; also *Rhet.,* 1356b 20–25 and *Top.* 157a 18–20.

[7] The lack of this definition in Aristotle's writings led W. D. Ross—one of the most eminent experts on Aristotle—to conclude that "the enthymeme is discussed in many passages of the *Rhetoric*, and it is impossible to extract from them a completely consistent theory of its nature" (Ross, 1949, p. 409).

[8] *Rhet.,* 1356a 22, b5; 57a 23; 94a 26; 95b 22; 00b 27 ff.; 02a 29 ff.

μὲν γὰρ ἀπόδειξις συλλογισμός τις), but not every deduction is a demonstration" (*APr,* 25b 29–31), it can be seen that the term "syllogism/deduction" is broader and contains more than strictly scientific (apodeictic) demonstration. It is a kind of deductive reasoning as long as it preserves the structure implied by its definition. Thus, syllogisms can occur not only in formally scientific argumentation, but also in dialectical or rhetorical argumentation (Grimaldi, 1972, p. 85).[9] In the *Rhetoric*, Aristotle states that "the enthymeme is a sort of deduction",[10] and claims that "he who is best able to see how and from what elements a deduction is produced will also be best skilled in the enthymeme" (*Rhet.*, 1355a 8–14).

On the basis of the theory of knowledge presented in the *Posterior Analytics*, it can be seen that the difference between a deduction in science and the enthymeme lies in the nature of premises assumed in a demonstrative and in a rhetorical deduction. In a scientific deduction, premises must be true, primitive, immediate, more familiar, prior to, and explanatory of, the conclusion,[11] whereas in the enthymeme they can be either probable or necessary (τεκμήρια). The probability of premises and conclusions indicates the affinity of rhetoric with dialectic, the syllogism of which is based on premises that are generally accepted (ἐξ ἐνδόξων; *Top.*, 100a 27–100b 18). The premises used in enthymematic reasoning, most of which are probable, do not exhaust the possibility of using the enthymeme. This means that a discourse in rhetoric can go beyond what is probable knowledge. From this it follows that a rhetorical syllogism, because of the nature of its premises (probable or necessary), may occur as a dialectical syllogism or, sometimes, as a strictly scientific (apodeictic)

[9] According to I. Hacking, who is worth quoting here, "It is widely agreed that *Topics* and *Rhetoric* represent some of Aristotle's first courses of lectures [...] *Topics* is about dialectic, back and forth argument between peers. Rhetoric is the argument of an orator addressing an audience. [...] This has a corollary which I shall call 'Before logic': Aristotle had not yet discovered the syllogism at the time he lectured on rhetoric and dialectic ... The syllogism introduced a new ritual into argument, one [a ritual] that was not simply there to discover [in the times of *Rhetoric* and *Topics*]. What was [radically new] was what we now call a valid form of argument. If the premises are true, then the conclusion must be true too. Aristotle, in creating the theory of the syllogism, discovered what we call logical consequence and valid argument" (Hacking, 2013, p. 426).

[10] In the original: ἐνθύμημα μὲν ῥητοπικὸν συλλογισμόν (*Rhet.,* 1356b 4–5).

[11] In the original: ἀληθῆ, πρῶτα καὶ ἄμεσα, γνωριμώτερα καὶ πρότερα καὶ αἴτια τοῦ συμπεράσματος (*APo,* 71b 19).

syllogism. Hence, the enthymeme seems to be a form of inference which may partake of both the nature of the dialectical and the scientific syllogism (Grimaldi, 1972, p. 86).

The question arises, however, about the formal construction of an enthymeme.[12] Aristotle's comments do not indicate that he considered the enthymeme to be an ordinary syllogism of three statements. Hence, a rhetorical syllogism has been commonly treated as a syllogism truncated in form, a syllogism with a suppressed premise or an omitted conclusion (Bitzer, 1959, p. 143).[13] Nevertheless, Aristotle's statements in the *Rhetoric* do not permit one—as it seems—to make this condition necessary when defining an enthymeme. Aristotle repeatedly pointed out that it was possible to omit a conclusion or leave out the major premise, but he did not treat this as the *sine qua non* condition for the enthymeme. The following passage from the *Rhetoric* can serve as an example:

> The enthymeme must consist of few propositions, fewer often than those which make up a primary deduction; For if any one of these propositions is a familiar fact, there is no need even to mention it, the hearer adds it himself. (*Rhet.,* 1357a 16–17)

These comments set a pragmatic condition for effective argumentation; namely we should not introduce premises that are unnecessary (from the point of view of the recipients), for instance, the premises that are obvious, as in the example with the winner at the Olympic games.[14] Aristotle's view on the form of an enthymeme is well summarised in his statement that enthymemes should be "as compact as possible" (*Rhet.,* 1419a 18–19); the enthymeme should be a brief, direct and condensed inference in the shortest possible form.[15]

---

[12] Bitzer's paper (1959) provides an overview of the main approaches to this problem.

[13] The enthymeme is treated in this way by Cope, Baldwin, and De Quincey, to name a few; and also in most textbooks on logic (cf. e.g. Lechniak, 2012, p. 212).

[14] "For instance, to prove that Dorieus was the victor in a contest at which the prize was a crown, it is enough to say that he won a victory at the Olympic games; there is no need to add that the prize at the Olympic games is a crown, for everybody knows it" (*Rhet.,* 1357a 18–21).

[15] Aristotle's exposition on maxims as a means of persuasion points to this as well. "Now an enthymeme is a deduction [...], it is therefore roughly true that the premisses or conclusions of enthymemes, considered apart from the rest of argument, are maxims" (Rhet., 1394a 26–28). A maxim is transformed into a full en-

This requirement, that enthymemes should be as much condensed as possible is determined by the factor which always plays a key role in a rhetorical speech, namely the presence of the audience. Aristotle, as Grimaldi notes (1972, p. 88), is concerned here that the auditors acquire the knowledge and understanding of the subject of a speech, an understanding that he calls μάθησις ταχεῖα (a quick, comprehensive grasp of the problem).[16] "A quick grasp of the problem"—as he writes in Book III of the *Rhetoric*—is achieved in three ways: 1) by enthymeme with respect to thought, 2) by antithesis with respect to style (antithetic style), and 3) in language by metaphor (*Rhet.*, 1410b 27–36). Thus, Aristotle focuses on three components of speech: thought, language, and style. The enthymeme does it by the way in which it organizes the thought; the clarity of style does it by the way in which the idea is emphasized by the sentence structure; in language, in turn, it is the structure of analogy in the metaphor, which results in "a quick grasp" (μάθησις ταχεῖα). The relation between enthymeme and antithetic style is emphasized by Aristotle's statement that "so too in enthymemes a compact and antithetical utterance passes for an enthymeme, such language being the proper province of enthymeme (χώρα ἐστιν ἐνθυμήματος)" (*Rhet.,* 1401a 4–6). The antithesis is based on the relation between two concepts or premises, thanks to which we can move directly from a concept that is known to a new one, or from a premise already known to a lesser known one. As Hacking points out, there is a fundamental practical difference between dialectic and rhetoric.

Rhetoric is concerned with discourse addressed to an audience and audiences have short attention spans. That is why, long arguments should be avoided. Because of this need for brevity, agreed common knowledge is always the best starting point. When the orator is familiar with the audience, most of the premises can be assumed, not stated. Dialectic, by contrast, is argument between two parties. It is back and forth. Steps can be recalled, repeated, defended, and criticized, collectively or one by one. Dialectic is dialogue. Rhetoric is monologue (Hacking, 2013, p. 429).

The stylistic construction of an utterance (antithetic style) and the form of an enthymeme (where one premise is omitted), focus above all on the simplicity and directness which are necessary for the audience to un-

---

thymeme when the reason or justification for a given statement that forms a premise or a conclusion, is added.

[16] *Rhet.,* 1410b, 10–12, 20–21, 25–26; 1400b, 31–34; 1357, 21.

derstand the utterance. Introducing a complete deduction into the theory of rhetoric, could prevent the audience from understanding the message or, at the very least, would make this understanding difficult. Thus, a proposition is omitted in the enthymeme because of some *praxis* and because it is obvious. It exists, yet it is not explicitly stated. In this sense, formally speaking, the enthymeme is a normal syllogism, but it differs from a dialectical and demonstrative syllogism in assumed premises, or in the way the statements implied in the conclusion are qualified.

2. Argumentation by Enthymeme

For Aristotle, the fundamental difference between different kinds of syllogisms lies in the type of knowledge that is obtained in the conclusion. "Now the materials of enthymemes are probabilities and signs, so that each of the former must be the same as one of these" (*Rhet.*, 1357a 32–33). This remark is complemented by the statement that "enthymemes are based upon one or other of four things: a) probabilities (εἰκός), b) examples (παράδειγμα), c) evidences (τεκμήριον), d) signs (σημεῖον)" (*Rhet.*, 1402b 12–14). These "four things", however, can be reduced to just two. An example may be a source of enthymeme insofar as it can give you, on the basis of similar cases, a probable universal principle or truth from which you may then argue by the use of enthymeme to a particular inference (*Rhet.*, 1402b 15–17). An example gives *the universal* by that flash of insight by which we pass from knowledge of a particular fact to direct knowledge of the corresponding principle (Grimaldi, 1972, p. 104). In this context, it should be viewed as the basis for educing a universal proposition or principle. Evidence, on the other hand, is in fact a kind of sign because "of signs, one kind bears the same relation as the particular bears to the universal, the other the same as the universal bears to the particular. A necessary sign is an evidence (τεκμήριον), a non-necessary sign has no specific name (ἀνώνιμον)."[17] So, we are left with an enthymeme that is based on probabilities (ἐξ εἰκότων) and an enthymeme that draws its premises from signs (ἐκ σημείων).

[17] *Rhet.,* 1357b, 1–7. Podbielski renders the term *tekmerion* (τεκμήριον) as "evidence" in the sense of a necessary sign; for example, the presence of milk is a necessary sign that a woman is pregnant or has recently borne a child, which should be distinguished from a probable sign (for instance, the paleness of a woman may indicate pregnancy, but not necessarily, because it may also be a symptom of something completely different).

The differences between these two types of enthymeme are pointed out in the *Prior Analytics*: "*eikos* and *semeion* are not identical, a probability is a reputable proposition (ἔνδοξος) [...], a sign is meant to be a demonstrative proposition, either necessary or reputable (πρότασις ἀποδικτικὴ ἀναγκαία ἢ ἔνδοξος)"(*APr*, 70a 3–8). The difference between these two sources is ultimately based on the kind of knowledge obtained when we use either *semeion* or *eikos*. An enthymeme built upon a probability (εἰκοτα)—as Grimaldi notes (Grimaldi, 1972, p. 105 ff)—will give what is called the *ratio essendi* of the fact stated in the conclusion, that is the explanation why this conclusion actually is. In other words, premises contain the reasons for the fact stated in the conclusion. On the other hand, an enthymeme built upon signs (σημεῖα) indicates the *ratio cognoscendi* of the fact stated in the conclusion; i.e., it indicates a symptom from which this fact can be inferred, as it is in the proof from signs in the first figure.

In order to get a good understanding of this distinction between *ratio essendi* and *ratio cognoscendi*, it is necessary to review Aristotle's theory of syllogism in more detail. Aristotle differentiated three syllogistic figures,[18] namely:

Figure I	Figure II	Figure III
B is A	B is A	C is A
C is B	C is A	C is B
C is A	C is B	B is A

The methodological function of each premise is determined by the function of terms in a syllogism.[19] When analysing the role of terms in a syllogism, we can distinguish their logical function and the function "from the thing". The first one refers to the place that a term takes in a given syllogism (especially when it comes to the middle term, which

[18] Figure IV, which combines the remaining generally valid syllogistic modes, was given by Galen. Obviously, the above diagram shows only how the terms are located in relation to one another—premises and a conclusion can be both universal and particular, affirmative and negative.

[19] Obviously, from the purely formal side, there is no difference between major and minor premises (as premises exist in conjunction, and this is alternating); the findings on the role of premises in a syllogism are based on Achmanow's explication (Achmanow, 1965, pp. 224–237).

appears in both premises), while the other is that of the ontological cause (reason) of what we state in the conclusion on the subject.[20] These functions are convergent only in syllogisms of the first figure and can be illustrated by the following table:

Functions "from the thing" in Figure I

Middle term	objective reason why something belongs (or does not belong) to the subject	that which is near	B
Major term	property attributed (or denied) to the subject on the basis of the reason from which it follows	does not twinkle	A
Minor term	the thing to which we attribute (or deny) something on the basis of knowledge about why something belongs (does not belong) to it	planets	C

The third column of the table refers to a well-known example given by Aristotle in Chapter 13 of Book I of the *Posterior Analytics*:

$$\frac{\begin{array}{c}\text{What is near (B) does not twinkle (A)}\\ \text{Planets (C) are near (B)}\end{array}}{\text{Planets (C) do not twinkle (A)}}$$

> The middle term corresponds to the cause of the property that is attributed to the subject in the conclusion, the conclusion follows from the premises not only from necessity, but also because it contains knowledge of a causal relationship, which as such is necessary, so it must be necessarily true [...] In this case, the major premise shows the cause and its consequences, and the minor premise indicates the presence of this cause in the subject of reasoning. (Achmanow, 1965, p. 228)

Consequently, this syllogism is an example of a syllogism based on the *ratio essendi*. However, as Aristotle notes, it is not always the case. He

[20] "All these [causes] are proved through the middle term. The case in which if something holds it is necessary that this does, does not occur if one proposition is assumed, but only if at least two are; and this occurs when they have one middle term. So when this one thing is assumed it is necessary for the conclusion to hold" (*APo*, 11, 94a 23–27).

gives the following example of a syllogism that is not based on knowing the cause (*APo*, 78a):

$$\frac{\begin{array}{c}\text{What does not twinkle (B) is near (A)}\\ \text{Planets (C) do not twinkle (B)}\end{array}}{\text{Planets (C) are near (A)}}$$

This syllogism is not from the knowledge of the reason why, but from the knowledge of what something is—planets are not near because they do not twinkle, but they do not twinkle because they are near.

> Although the conclusion necessarily follows from the premisses, it can not be considered to be necessarily true, because the fact that some subject is attributed with the consequence of some property does not make it necessary for the subject to possess that property itself. (Achmanow, 1965, p. 228)

What we have here is an example of a syllogism in *modus cognoscendi*. In a syllogism based on knowing the cause, logical motivation corresponds to the real cause of some property—that is why we have both the necessity of following and the necessity of a real presence of some property in the subject; this is not the case in a syllogism that is not based on the knowledge of the cause—"logical motivation does not correspond to the real cause of this property" (Achmanow, 1965, p. 228).

The definition of probability in the *Rhetoric* helps get a better understanding of *eikos* argumentation:

> a probability is a thing that happens for the most part—not, however, as some definitions would suggest, anything whatever that so happens, but only if it belongs to the class of what can turn out otherwise, and bears the same relation to that in respect of which it is probable as the universal bears to the particular. (*Rhet.*, 1357a 34b 1)

Probability is based on the typicality and regularity of some properties attributed to a given class of things, and the fact that some property is attributed is a condition for inference. A premise must be known and generally accepted.[21] Accepting the premises based on *eikos* leads to fur-

[21] As D. Walton (2001) points out, when talking about eikos, it would be better to use the word plausibility instead of probability.

ther knowledge that meets the condition of logicality on the one hand (as the conclusion implied by these premises is based on the rules of inference), and, on the other hand, these premises are acceptable to the mind because what they state corresponds to the observed facts, which is a condition for the mind to think that such is the actual fact. *Eikos* expresses an aspect of the real order that is understandable and stable. An inference from *eikos* does not conclude to an unconditioned and necessary truth; but it does present an eminently reasonable guaranty that the conclusion represents the objective fact (Grimaldi, 1972, p. 109 ff).

On the other hand, when writing about a sign in the *Prior Analytics* (*APr*, 70a, 7–9), Aristotle points to a relationship between two realities in the order of existence, which leads from the knowledge of one to the knowledge of the other. A sign is a relation between "two things" which have their foundation in the nature of these realities and their existence is objective and determined only by the fact that the existence of one depends on the existence of the other. The relationship between the sign and the signate leads the mind from the known to the unknown because of this one-to-one correspondence. It is a real relationship which has its ground in the *esse* of the sign and as such it is the relationship of formal causality (Grimaldi, 1972, p. 110). Because of the sign, we can know the signate. That is why, Aristotle believes that *semeion* has a stronger demonstrative force than *eikos*. This can be easily seen in Chapter 27 of the *Prior Analytics*, where he discusses the use of a sign in syllogistic figures. In general, the demonstrative force of a sign is expressed by the statement that "a sign wants to be a demonstrative proposition either necessary or reputable."[22] What follows is that there are different kinds of signs: necessary and commonly accepted (ἢ ἀναγκαῖα ἢ ἔνδοξος), which seems to correspond with the distinction made in the *Rhetoric* between necessary signs (τεκμήριον) and non-necessary signs (σημεῖον ἀνώνιμον). *Tekmerion* contains within itself an element of necessity in relation to the signate (πρότασις ἀποδεικτικὴ ἀναγκαῖα), while *semeion anonymon* indicates the signate only with probability (πρότασις ἀποδεικτικὴ ἔνδοξος). This distinction can be seen in the position of terms in a syllogism. *Tekmerion* is the middle term of an enthymeme or of a syllogism of the first figure and assumes the relation of necessity in respect to the signate.

This is the case in enthymemes of the first figure. We have:

[22] In the original: σημεῖον δὲ βούλεται εἶναι πρότασις ἀποδεικτική (*APr*, 70a 6–7).

[Every woman who has milk (B) is with child (A)]
This woman (C) has milk (B)

This woman (C) is with child (A)

If we state only the second premise—we have a sign; but if the first (implicit) premise is stated as well—we get a syllogism (deduction). As can be seen, such a rhetorical syllogism is a syllogism "from the thing", as the logical function of the middle term coincides with its "causal" function.

Semeion anonymon is the extreme term of inference and does not signify necessity. In turn, *semeion anonymon* as the middle term is identified in the second and third figure. "[Deduction] which proceeds through the last figure is refutable even if the conclusion is true, since the deduction is not universal nor relevant to the matter in question." On the other hand,

> the deduction which proceeds through the middle figure (II) is always refutable in any case; for a deduction can never be formed when the terms are related in this way; for though a woman with child is pale, and this woman is pale, it is not necessary that she should be with child. (*APr*, 70a 30–37)

For Aristotle's second figure, the example can be represented as follows (symbols A, B, C refer to symbols from "the thing"):

A woman with child (B) is pale (A)
This woman (C) is pale (A)

This woman (C) is with child (B)

The argumentation aims to prove that a woman is pregnant, and the reason is paleness as something that accompanies pregnancy and can be stated about the woman; if there is only the second premise, we have a sign; if both premises occur together, we get a syllogism. "In the enthymeme reduced to the second figure, the sign (paleness) is the middle term when we consider its logical function, but due to its nature (as a consequence) it should be called the major term and denoted by letter A" (Achmanow, 1965, p. 319). The situation is similar with enthymemes of the third figure.

Figure II

Aim:	to conclude that there is no objective reason in the subject on the basis of the lack of consequence in the subject.
Formal effect:	both premises cannot be affirmative
Major premise:	universal: expresses the relationship of cause and effect.
Negative consequence in the major premise:	minor premise: attributes the opposite to the subject—it is affirmative.
Affirmative consequence in the major premise:	minor premise contradicts the occurrence of the consequence in the subject.
Conclusion for the enthymeme:	sign in the second figure—consequence; is not a demonstrative sign.

3. The Apparent Enthymeme

As we have shown above, enthymemes from signs and from probabilities can quite easily be reduced to a demonstrative syllogism and as such can be examined by means of "ordinary" methods that are used to determine whether a deduction is valid (they differ from a demonstrative syllogism only in the kind of premises). Things are different when it comes to the apparent syllogism.[23] Aristotle's exposition on the *apparent* enthymeme and the *refutative* enthymeme serves to:

(a) reveal possible errors and evasions in logical reasoning;

(b) show how to contend with them. This is the defence of the logos against misleading and incorrect argument.

[23] Grimaldi notes that "there is rarely any discussion of what Aristotle calls the apparent enthymeme and the refutative enthymeme. The reticence is surprising since they represent another aspect of the enthymeme and an understanding of them would seem necessary to a full comprehension of enthymeme and enthymematic reasoning. In the present context they are particularly relevant and instructive for they confirm the three points just mentioned in the discussion of the enthymeme as the instrument of deductive reasoning: 1) the fact that rhetoric is concerned with truth, 2) the structural form of the enthymeme, and, 3) the character of its subject-matter" (Grimaldi, 1972, p. 94).

In the *Sophistical Refutations*[24] and the *Rhetoric* (B24), Aristotle classifies nine topoi as examples for the apparent enthymeme, which is considered to be specious reasoning, i.e. reasoning that is logically invalid. These specious inferences, can be divided into three groups:

(i) formally fallacious—treated as a syllogism in one of the three syllogistic figures, they contain a formal error;

(ii) materially fallacious—the content of statements (premises) of such an enthymeme is false—unnecessary, unlikely, or impossible.

(iii) inference that combines some lack in the syllogistic form and in reasoning from seemingly plausible premises, and thus imitates inference, which in fact does not take place, for example: "[...] some he saved, others he avenged, the Greeks he freed" (*Rhet.,* 1401a 10–11); each of these statements has been proved on the basis of other premises or arguments.

Ad (I) Enthymemes of the first group in the catalogue from the *Rhetoric* [B24] include topoi Ib, II, VIII, IX. These inferences are formally incorrect, namely:

- Ib follows from the use of homonymy to give the appearance of inference;
- II takes the whole and its parts as identical, though often they are not;[25]
- VIII—fallacy lies in omitting the middle term;

[24] The *Sophistical Refutations* (165b 23 ff) give two kinds of "false" inference: (i) παρὰ τὴν λέξειν (*fallacia dictionis*)—inference based on the use of linguistic forms that "seem to refute a statement"; apparent deductions make use of the following linguistic forms: 1) homonymy (ὁμωνυμία), 2) amphiboly—ambiguous words (ἀμφιβολία), 3) combination of expressions (συνθέσις), 4) division of expressions (διαίρεσις), 5) prosody, or changing the length of vowels (προσῳδία), 6) incorrect grammatical forms (σχῆμα λέξεως). (ii) ἔξω τῆς λέξεως (*fallacia extra dicionem*)—inference based on the erroneous use of non-linguistic forms.

[25] Fallacy of the statement: "The one who knows the letters knows the whole word, since the word is the same thing as the letters which compose it", can be demonstrated by the following reconstruction: Who knows [all] parts of the whole, knows the whole. Each word is a whole made up of letters. Hence, anyone who knows all the letters [that make up a word] knows this word (*Rhet.*, 1401a 28–29).

- IX (*fallacia secundum quid est similiter*)—is based on using an expression in an absolute sense (i.e. without qualification) and in a particular sense interchangeably;[26]

Ad (ii) Materially fallacious inference, where fallacy of one of the premises can be caused by topoi V–VII, which are formally fallacious:[27]

- V (*fallacia accidentis*; *Soph. Ref.*, 166b 28–32)—fallacy that occurs because it is assumed that the same applies to a thing as to one of its attributes;
- VI (*fallacia consequentis*; *Soph. Ref.*, 167b 1–9)—fallacy stems from the belief that the relation of consequence is convertible; i.e., if we assume that every A is B, then every B is also A (or in other words, by assuming that if there is A, then there is B, it is assumed that if there is B, there is also A);
- VII (*fallacia propter non causam ut causa*; *Soph. Ref.*, 167b 21 ff)—accepts the principle that because an event happened earlier, it is a cause of a later event (*post hoc ergo propter hoc*).

Ad (iii) Inference that is fallacious both because of the form of a syllogism and of its content: by using seemingly probable premises (IV) or by suggesting that they follow from some reasoning that, in fact, is missing (Ia, III).

An important property of the apparent enthymeme is that it inadequately represents reality as it is and as it can be known (Grimaldi, 1972, p. 95), because "what makes a sophist is not his abilities but his choices"

[26] Reconstruction of an example: What is not is an object of opinion. Whatever is an object of opinion is [as an object of opinion]. Therefore, what is not, is [as an object of opinion]. Normally, taking into account the information in square brackets, we have the Barbara syllogism; but deleting the information in the brackets changes the relative meaning into the absolute one. Then we have a distinction: "is (in reality)"—"is (as an object of opinion)" (*Soph. Ref.*, 166b 37–167a 19).

[27] Strictly speaking, topoi V–VII, just as the topoi of group I, are also examples of formally fallacious inferences. What makes them different from the topoi of group I is that they are used as an apparent proof for premises (and not as a proof for the conclusion, as is the case in group I). They result in false premises. More properly, we would say that the premises in the topoi of group II are fallaciously justified (*petitio principii*).

(*Rhet.,* 1355b 17–18). In all cases, the apparent enthymeme does not validly demonstrate the probable knowledge; i.e., the knowledge concerning the contingent reality, but it usually gives the appearance of demonstrating—φαίνεσθαι δεικνύναι (*Rhet.,* 1356a 36). Aristotle also uses the term "eristic syllogism", or "eristic (contentious) deduction" for the apparent enthymeme (*Top.* 100B 13–101a 4),[28] and by that he understands those arguments "that deduce or appear to deduce to a conclusion from premises that appear to be reputable but are not so" (*Soph. Ref.,* 165b 7–8).

4. The Refutative Enthymeme

According to Aristotle, an argument may be refuted in two ways: 1) by a counter-deduction (ἀντισυλλογισάμενον), or 2) by bringing an objection (ἔνστασιν) (*Rhet.*, 1402a 31).

Ad (1) The difference between the demonstrative (deictic) enthymeme and the refutative enthymeme (elenctic) is determined by placing logical argumentation in rhetoric into the context of dialectical argumentation:

> [...] there are two kinds of enthymemes. One kind proves some affirmative or negative proposition; the other kind disproves one. The difference between the two kinds is the same as that between refutation and deduction in dialectic. The probative enthymeme makes an inference from what is accepted, the refutative makes an inference to what is unaccepted. (*Rhet.,* 1396b 23–28)

Thus, the relation between deictic and elenctic enthymeme in rhetoric is analogous to the relation between a dialectical syllogism and elenchos in dialectics (*Soph. Ref.*, 164b 27–165a 3). "As elenchos and the dialectical syllogism are both syllogisms, one destructive, the other constructive, so are the elenctic and deictic enthymemes both enthymemes. Any difference between them resides solely in the fact that the elenctic enthymeme (just as elenchos itself) is inference directed to disprove the conclusion reached

[28] According to Aristotle, there are three types of reasoning depending on the purpose and nature /content of premises: (1) scientific reasoning/reasoning used in science—aimed at reaching the truth; and proceeding from true / necessary premises; (2) reasoning in rhetoric—aimed at defeating an opponent; here premises are probable, i.e. believed by most people—ἐξ ἐνδόξων; (3) eristic / sophistical reasoning—the content of a dispute is not important; this kind of dispute called γωνικῶς or ἐριστικῶς was practised by Sophists, and it is the subject of Aristotle's *Sophistical Refutations.*

by the deictic enthymeme that it is refuting (Grimaldi, 1972, p. 100)."[29] Deictic and elenctic enthymemes use the same topoi and these topoi, categories of reasoning, are usually based on probabilities (ἐκ τῶν ἐνδόξων), which results in the fact that many of them are contradictory to one another (*Rhet.,* 1402a 33–35). Since opposing probabilities are possible, there is a reason for using the refutative enthymeme in order to infer a conclusion that negates the conclusion of a demonstrative enthymeme while keeping the same categories of argument.

Ad (2) "An objection (ἔνστασις) is a proposition contrary to a proposition" (*APr*, 69a 37); *enstasis* consists in standing in the way of an opponent's reasoning by denying one of his premises, before he formulates a syllogism which should be answered with a counter-syllogism. *Enstasis* questions universal premises and it must be made in the same figure in which the initial syllogism was formulated (Aristotle, Polish ed. 1990, p. 247, note 95).

In the *Rhetoric*, Aristotle gives four ways of raising objections to an opponent's premises: "Objections, as appears in the *Topics*, may be raised in four ways—either by directly attacking your opponent's own statement, or by putting forward another statement like it, or by putting forward a statement contrary to it, or by quoting previous decisions."[30] In his commentary to the *Prior Analytics*, Kazimierz Leśniak gives a brief and clear explanation of these four ways. An objection (ἔνστασις) can be raised:

1) on the basis of the thing itself (ἐξ ἑαυτοῦ)—if someone claims that love is good, we object either a) by stating that every need is bad, which is a universal statement, or b) by stating that unhappy love is bad, which is a particular statement.[31]

2) on the basis of a similarity (ἐκ τοῦ ὁμοίου)—if a statement that we question says that those who have been badly treated hate those who

[29] Cf. *Rhet.*, 1403a 15–31, also 1418b 2–6.

[30] αἱ δ' ἐνστάσεις φέρονται καθάπερ καὶ ἐν τοῖς τοπικοῖς τετραχῶς ἢ γὰρ ἐξ ἑαυτοῦ ἢ ἐκ τοῦ ὁμοίου ἢ ἐκ τοῦ ἐναντίου ἢ ἐκ τῶν κεκιμμένων (*Rhet.*, 1402a 34 ff).

[31] Aristotle's initial argument can be presented in the form of reasoning: Every need to do good is good (P; enthymematic premise). Love is the need to do good (Q). Therefore, every love is good (R). Using the first method, we refute the major premise with the argument: Every lack is evil. Every need is a lack. Therefore, every need is evil. Therefore, the need to do good, is evil.

treated them badly, we reply that those who have been well treated do not always treat well those who treated them well.[32]

3) on the basis of a contradiction (ἐκ τοῦ ἐναντίου)—if someone claims that a good person does good to all his friends, we reply that a bad person does not do evil to all his friends.

4) on the basis of previous decisions (ἐκ τῶν κεκιμμένων)—if the statement that we question says that we should always be forgiving to drunken people, we reply that Pittakos is by no means worthy of praise, because if he were he would not deserve stricter punishment than the one who being drunk did bad things (Aristotle, Polish ed. 1990, p. 248, note 99).

From what has been written above, it can be concluded that *enstasis* is a probable proposition that suggests that an opponent has made a false statement, or strictly speaking, that undermines his belief in the truth of the claim he has made by challenging one of his premises or showing that his reasoning to justify the premise is invalid. This explanation corresponds to the definition of *enstasis* given in the *Prior Analytics*, namely that "*enstasis* is a proposition contrary to a proposition" (*APr*, 69a 37). The use of *enstasis* in challenging an argument can be considered from the perspective of contemporary non-classical logics. The classical propositional calculus (and classical consequence) fails to provide an adequate view of argumentation by *enstasis*. The core of this argumentation is to "block" an opponent's argument by challenging his premise. Meanwhile, classical logic is monotonic; i.e.: If $X \vdash \varphi$, then $(X \cup \psi) \vdash \varphi$ (if premises are contradictory, then a set of propositions derived from them is contradictory and hence trivial). Thus, adding the *enstasis* to premise, will lead the system of conclusions into collapse (contradiction). From the point of view of the theory of argumentation, such an approach to blocking

[32] Here again, the challenged argument can be presented in the form of the Barbara syllogism: Everyone who has suffered distress, hates. Everyone who has suffered evil, has suffered distress. Therefore, everyone who has suffered evil, hates. The first premise of this argument can be challenged by means of an antithesis: "Those who have experienced good, do not always love." This antithesis can be supported by an argument: [Each] experience of good is similar to the experience of evil. Some who experience good do not love. Therefore, some who suffer evil do not hate.

a premise is obviously undesirable. It seems that non-monotonic logics, for example, can be a useful tool here.[33]

> We are said to be reasoning non-monotonically when we allow that a conclusion that is well drawn from given information may need to be withdrawn when we come into possession of further information, even when none of the old premises is abandoned. In brief, a consequence relation is non-monotonic iff it can happen that a proposition x is a consequence of a set A of propositions, but not a consequence of some superset A ∪ B of A. (Makinson, 2008, p. 2)

To come back, for example,[34] to the *enstasis* on the basis of the thing itself (ἐξ ἑαυτοῦ) ("if someone claims that love is good, we object either a) by stating that every need is bad, which is a universal statement, or b) by stating that unhappy love is bad, which is a particular statement"): the thesis that love is good is based on implied assumptions—*enstasis*

[33] Formal theories of belief revision can serve as another tool here. They describe formal conditions for rational revision of beliefs; that is, adding (expanding), removing (contracting) and "exchanging" a given belief into a belief that contradicts it (revision). The operation of contracting would be the closest to *enstasis*: an argument that we give forces the opponent to give up his belief about the truth of a premise initially accepted. For more details on the formal theory of belief revision, see (Lechniak, 2011). On the other hand, in the so-called formal epistemology, there is the concept of defeasible reasoning developed by J. Pollock. What is essential in this theory is the distinction made between defeasible schemes and indefeasible schemes. Reasoning in line with defeasible schemes provides reasons for a conclusion and mandates a conclusion if there is no information that would contradict this conclusion. A set of defeaters that may challenge the justification of the conclusion is associated with the schemes of defeasible reasoning. Reasoning is indefeasible if a set of defeaters is not associated with it (e.g. reasoning based on the laws of logic). Two kinds of defeaters can be distinguished: the rebutting defeater, which is an argument for the opposite conclusion (any reason for denying the conclusion), and the undercutting defeater, which attacks the inference between the premises and the conclusion of defeasible reasoning; cf. (Pollock, 2008) and /or (Pollock & Gillies, 2000). As a reviewer of this article rightly suggests, rebutting defeaters can be related to the issue of contradictory syllogisms, and undercutting defeaters—to using topoi based on fallible, in some cases, forms of inference.

[34] The above attempt is only preliminary and there is no doubt that it requires refining; our aim is just to show that *enstasis* can be described in the language of non-monotonic logics.

attacks the implied premise that every need to do good is good. Using the sign $\models$ for the enthymematic inference,[35] we can write the initial reasoning that is attacked as $P \wedge Q \models R$, while the counter-argument ("Every need is evil" (S)) added to a set of premises negates the conclusion; i.e., $(P \wedge Q \wedge S) \models \neg R$, and consequently $(P \wedge Q \wedge S) \not\models R$.

5. Summary

In summary, the following conclusions can be drawn:

(i) the enthymeme that proceeds from what is probable (εἶκος) and from what is necessary (σημεῖον ἀνώνιμον) implies conclusions corresponding to its suppositions; that is why, conclusions can be only probable in a (rhetorical) syllogism, or they can be strictly scientific statements (τεκμήριον), as is the case with conclusions in an apodeictic syllogism.

(ii) demonstrative and refutative enthymemes do not differ (taking into account the omitted major premise) in their structure from apodeictic syllogisms; the difference lies in their premises. Since the aim of an enthymeme is rhetorical (to convince the listener), the argument must be concise and that is why the major premise is omitted (as the implied one).

(iii) the conciseness of an enthymeme makes it possible to use apparent enthymemes, i.e. reasoning that is logically invalid; when such an apparent enthymeme is "expanded" into a full syllogism, this invalidity becomes obvious.

(iv) contemporary non-monotonic logics (e.g. default logic, defeasible logic or the theory of belief revision) can be useful in the analysis of enthymematic argumentation.[36]

---

[35] J. Malinowski (1997) points out that, just as in the classical formalization of reasoning, we would use the following statements: "If P is true, then Q must be true" or "If we accept P, then we must accept Q", so in the formalization of common reasoning we would use statements such as "If P, then it is usually Q", "If P is acceptable, then Q is acceptable", "If P is probable, then Q is probable."

[36] To date, we have not found any studies that would show how these logics can be practically applied in the formal analysis of an enthymeme.

REFERENCES

Achmanow, A. (1965). *Logika Arystotelesa*. Warsaw: PWN.

Bitzer, L. (1959). Aristotle's Enthymeme Revisited. *Quarterly Journal of Speech*, *45*(4), 399–408. Reprinted in: K. V. Erickson (Ed.), *Classical Heritage of Rhetoric* (pp. 141–155). Metuchen, New Jersey: Scarecrow Press.

Grimaldi, W. M. A. (1972). *Studies in the Philosophy of Aristotle's Rhetoric*. Wiesbaden: Franz Steiner Verlag.

Hacking, I. (2013). What Logic Did to Rhetoric. *Journal of Cognition and Culture, 13*(5), 419–436.

Lechniak, M. (2011). *Przekonanie i zmiana przekonań*. Lublin: Wyd. KUL.

Lechniak, M. (2012). *Elementy logiki dla prawników*. Lublin: Wyd. KUL.

Malinowski, J. (1997). Logika niemonotoniczna. *Przegląd Filozoficzny. Nowa Seria*, *21*(1), 37–53.

Madden, E. H. (1952). The Enthymeme: Crossroads of Logic, Rhetoric and Metaphysics. *Philosophical Review*, *61*(3), 368–376.

Madden, E. H. (1957). Aristotle's Treatment of Probability and Signs. *Philosophy of Science*, *24*(2), 167–172.

Makinson, D. (2008). *Od logiki klasycznej do niemonotonicznej*. Toruń: Wyd. Naukowe UMK.

Malinowski, J. (1997). Logika niemonotoniczna. *Przegląd Filozoficzny. Nowa Seria*, *21*(1), 37–53.

McBuruney, J. H. (1936). The Place of Enthymeme in Rhetorical Theory. *Speech Monographs*, *3*, 49–74. Reprinted in: K. V. Erickson (Ed.), *Aristotle, The Classical Heritage of Rhetoric* (pp. 117–140). Metuchen, New Jersey: Scarecrow Press.

Perelman, Ch. (2002). *Imperium retoryki*. Warsaw: PWN.

Pollock, J. (2008). *Defeasible Reasoning*. In: J. Adler, L. Rips (Eds.), *Reasoning: Studies of Human Inference and its Foundations* (pp. 451–470). Cambridge: Cambridge University Press.

Pollock, J., Gillies, A. (2000). Belief Revision and Epistemology. *Synthese, 122*(1–2), 69–92.

Ross, W. D. (1949). *Aristotle's Prior and Posterior Analytics*. Oxford: Oxford University Press.

Sprute, J. (1982). *Die Enthymemtheorie der aristotelischen Rhetorik*. Gottingen: Vandenhoeck & Ruprecht.

Walton, D. (2001). Enthymemes, Common Knowledge, and Plausible Inference. *Philosophy and Rhetoric*, *34*(2), 93–112.

Originally published as "Strategie argumentacji w teorii retoryki Arystotelesa: entymematy pozorne i obalające". *Studia Semiotyczne*, *32*(1), 61–82, DOI: 10.26333/sts.xxxii1.04. Translated by Marta Cechowicz.

ABOUT THE AUTHORS

Gabriela Besler, dr hab., adiunkt UŚ, Instytut Filozofii, Wydział Nauk Społecznych, ul. Bankowa 11, 40-007 Katowice. ORCID: 0000-0002-1843-5198.

Marek Lechniak, dr hab., prof. KUL, Katedra Logiki, Wydział Filozofii, 20-950 Lublin, al. Racławickie 14. ORCID: 0000-0002-0768-7963.

Maciej Sendłak, dr, adiunkt, Instytut Filozofii, Zakład Filozofii Analitycznej, Uniwersytet Warszawski, Krakowskie Przedmieście 3, 00-927 Warszawa. ORCID: 0000-0002-0539-5924.

Paweł Stacewicz, dr inż, adiunkt PW, Wydział Administracji i Nauk Społecznych, Plac Politechniki 1, 00-661 Warszawa. ORCID: 0000-0003-2500-4086.

Andrzej Stefańczyk, dr, adiunkt KUL, Katedra Historii Filozofii Sta-rożytnej i Średniowiecznej, Wydział Filozofii, 20-950 Lublin, al. Racławickie 14. ORCID: 0000-0001-5621-0777.

Jacek Wawer, dr, adiunkt, Instytut Filozofii, Zakład Epistemologii, Uniwersytet Jagielloński, ul. Grodzka 52, 31-044 Kraków. ORCID: 0000-0003-2546-0962.

Piotr Wilkin, dr, Instytut Filozofii, Uniwersytet Warszawski. ORCID: 0000-0003-4714-5269.

Krzysztof Wójtowicz, prof. dr hab., prof. UW, Instytut Filozofii, Wydział Filozofii i Socjologii, Krakowskie Przedmieście 3, 00-927 Warszawa. ORCID: 0000-0002-1187-8762.